

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
20 December 2001 (20.12.2001)

PCT

(10) International Publication Number
WO 01/96551 A2

- (51) International Patent Classification⁷: C12N 15/00 [US/US]; 6801 Paseo Delicias, P.O. Box 7214, Rancho Santa Fe, CA 92067-7214 (US).
- (21) International Application Number: PCT/US01/19367
- (22) International Filing Date: 14 June 2001 (14.06.2001)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
09/594,459 14 June 2000 (14.06.2000) US
09/677,584 30 September 2000 (30.09.2000) US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:
US 09/594,459 (CIP)
Filed on 14 June 2000 (14.06.2000)
US 09/677,584 (CIP)
Filed on 30 September 2000 (30.09.2000)
- (71) Applicant (for all designated States except US): DIVERSA CORPORATION [US/US]; 4955 Directors Place, San Diego, CA 92121 (US).
- (72) Inventor; and
- (75) Inventor/Applicant (for US only): SHORT, Jay, M.
- (74) Agent: HAILE, Lisa, A.; Gray Cary Ware & Freidenrich LLP, Suite 1600, 4365 Executive Drive, San Diego, CA 92121-2189 (US).
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:
— without international search report and to be republished upon receipt of that report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/96551 A2

(54) Title: WHOLE CELL ENGINEERING BY MUTAGENIZING A SUBSTANTIAL PORTION OF A STARTING GENOME. COMBINING MUTATIONS, AND OPTIONALLY REPEATING

(57) Abstract: An invention comprising cellular transformation, directed evolution, and screening methods for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable. Also, a method of retooling genes and gene pathways by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, thus conferring operability to a novel gene pathway when it is introduced into an intended host. For example a novel man-made gene pathway, generated based on microbially-derived progenitor templates, that is operable in a plant cell. Furthermore, a method of generating novel host organisms having increased expression of desirable traits, recombinant genes, and gene products.

**WHOLE CELL ENGINEERING
BY MUTAGENIZING A SUBSTANTIAL PORTION OF A STARTING GENOME,
COMBINING MUTATIONS,
AND OPTIONALLY REPEATING**

A - FIELD OF THE INVENTION

This invention relates to the field of cellular and whole organism engineering. Specifically, this invention relates to a cellular transformation, directed evolution, and screening method for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable.

This invention also relates to the field of protein engineering. Specifically, this invention relates to a directed evolution method for preparing a polynucleotide encoding a polypeptide. More specifically, this invention relates to a method of using mutagenesis to generate a novel polynucleotide encoding a novel polypeptide, which novel polypeptide is itself an improved biological molecule &/or contributes to the generation of another improved biological molecule. More specifically still, this invention relates to a method of performing both non-stochastic polynucleotide chimerization and non-stochastic site-directed point mutagenesis.

Thus, in one aspect, this invention relates to a method of generating a progeny set of chimeric polynucleotide(s) by means that are synthetic and non-stochastic, and where the design of the progeny polynucleotide(s) is derived by analysis of a parental set of polynucleotides &/or of the polypeptides correspondingly encoded by the parental polynucleotides. In another aspect this invention relates to a method of performing site-directed mutagenesis using means that are exhaustive, systematic, and non-stochastic.

Furthermore this invention relates to a step of selecting from among a generated set of progeny molecules a subset comprised of particularly desirable species, including by a process termed end-selection, which subset may then be screened further. This invention also relates to the step of screening a set of polynucleotides for the production of a polypeptide &/or of another expressed biological molecule having a useful property.

Novel biological molecules whose manufacture is taught by this invention include genes, gene pathways, and any molecules whose expression is affected thereby, including directly encoded polypeptides &/or any molecules affected by such polypeptides. Said novel biological molecules include those that contain a carbohydrate, a lipid, a nucleic acid, &/or a protein component, and specific but non-limiting examples of these include antibiotics, antibodies, enzymes, and steroidal and non-steroidal hormones.

In a particular non-limiting aspect, the present invention relates to enzymes, particularly to thermostable enzymes, and to their generation by directed evolution. More particularly, the present invention relates to thermostable enzymes which are stable at high temperatures and which have improved activity at lower temperatures.

B - BACKGROUND

General Overview of the Problem to Be Solved

Brief Summary: It is instantly appreciated that the process of performing a genetic manipulation on a organism to achieve a genetic alteration, whether it is on a unicellular or on a multi-cellular organism, can lead to harmful, toxic, noxious, or even lethal effects on the manipulated organism. This is particularly true when the genetic manipulation becomes sizable. From a technical point of view, this problem is seen as one of the current obstacles that hinder the creation of genetically altered organisms having a large number of transgenic traits.

On the marketing side, is instantly appreciated that the purchase price of a genetically altered organism is often dictated by, or proportional to, the number of transgenic traits that have been introduced into the organism. Consequently, a genetically altered organism having a large number of stacked transgenic traits can be quite costly to produce and purchase and economically in low demand.

On the other hand, the generation of organism having but a single genetically introduced trait can also lead to the incurrence of undesirable costs, although for other reasons. It is thus appreciated that the separate production, marketing, & storage of genetically altered organisms each having a single transgenic traits can incur costs, including inventory costs, that are undesirable. For example, the storage of such organisms may require a separate bin to be used for each trait. Furthermore, the value of an organisms having a single particular trait is often intimately tied to the marketability of that particular

trait, and when that marketability diminishes, inventories of such organisms cannot be sold in other markets.

The instant invention solves these and other problems by providing a method of producing genetically altered organisms having a large number of stacked traits that are differentially activatable. Upon purchasing such a genetically altered organism (having a large number of differentially activatable stacked traits), the purchasing customer has the option of selecting and paying for particular traits among the total that can then be activated differentially. One economic advantage provided by this invention is that the storage of such genetically altered organisms is simplified since, for example, one bin could be used to store a large number of traits. Moreover, a single organism of this type can satisfy the demands for a variety of traits; consequently, such an organism can be sold in a variety of markets.

To achieve the production of genetically altered organisms having a large number of stacked traits that are differentially activatable, this invention provides - in one specific aspect - a process comprising the step of monitoring a cell or organism at holistic level. This serves as a way of collecting holistic - rather than isolated - information about a working cell or organism that is being subjected to a substantial amount of genetic manipulation. This invention further provides that this type of holistic monitoring can include the detection of all morphological, behavioral, and physical parameters.

Accordingly, the holistic monitoring provided by this invention can include the identification &/or quantification of all the genetic material contained in a working cell or organism (e.g. all nucleic acids including the entire genome, messenger RNA's, tRNA's, rRNA's, and mitochondrial nucleic acids, plasmids, phages, phagemids, viruses, as well as all episomal nucleic acids and endosymbiont nucleic acids). Furthermore this invention provides that this type of holistic monitoring can include all gene products produced by the working cell or organisms.

Furthermore, the holistic monitoring provided by this invention can include the identification &/or quantification of all molecules that are chemically at least in part protein in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part carbohydrate in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part proteoglycan in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification

of all molecules that are chemically at least in part glycoprotein in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part nucleic acids in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part lipids in a working cell or organism.

In one aspect, this invention provides that the ability to differentially activate a trait from among many, such as an enzyme from among many enzymes, depends on the enzyme(s) to be activated having a unique activity profile (or activity fingerprint). An enzyme's activity profile includes the reaction(s) it catalyzes and its specificity. Thus, an enzyme's activity profile includes its:

- Catalyzed reaction(s)
- Reaction type
- Natural substrate(s)
- Substrate spectrum
- Product spectrum
- Inhibitor(s)
- Cofactor(s)/prosthetic group(s)
- Metal compounds/salts that affect it
- Turnover number
- Specific activity
- K_m value
- pH optimum
- pH range
- Temperature optimum
- Temperature range

It is also instantly appreciated that enzymes are differentially affected by exposure to varying degrees of processing (e.g. upon extraction &/or purification) and exposure (e.g. to suboptimal storage conditions). Accordingly, enzyme differences may surface after exposure to:

- Isolation/Preparation
- Purification
- Crystallization
- Renaturation

It is instantly appreciated that differences in molecular stability can also be used advantageously to differentially activate or inactivate selected enzymes, by exposing the enzymes for an appropriate time to variations in:

- pH
- Temperature
- Oxidation
- Organic solvent(s)
- Miscellaneous storage conditions

It is thus appreciated that in order to be able to differentially activate selected traits among a plurality of stacked traits, it is desirable to introduce into a working cell or organism traits conferred by molecules (e.g. enzymes) having very unique profiles (e.g. unique enzyme fingerprints). Furthermore, it is appreciated that in order to obtain the molecules having a representation of a wide range of molecular fingerprints, it is advantageous to harvest molecules from the widest possible reaches nature's diversity. Thus, it is beneficial to harvest molecules not only from cultured mesophilic organisms, but also from extremophiles that are largely uncultured.

In another aspect, it is instantly appreciated that harvesting the full potential of nature's diversity can include both the step of discovery and the step of optimizing what is discovered. For example, the step of discovery allows one to mine biological molecules that have commercial utility. It is instantly appreciated that the ability to harvest the full richness of biodiversity, i.e. to mine biological molecules from a wide range of environmental conditions, is critical to the ability to discover novel molecules adapted to function under a wide variety of conditions, including extremes of conditions, such as may be found in a commercial application.

However, it is also instantly appreciated that only occasionally are there criteria for selection &/or survival in nature that point in the exact direction of particular commercial needs. Instead, it is often the case that a naturally occurring molecule will require a certain amount of change – from fine tuning to sweeping modification – in order to fulfill a particular unmet commercial need. Thus, to meet certain commercial needs (e.g., a need for a molecule that is functional under a specific set of commercial processing conditions) it is sometimes advantageous to experimentally modify a naturally expressed molecule to achieve properties beyond what natural evolution has provided &/or is likely to provide in the near future.

The approach, termed directed evolution, of experimentally modifying a biological molecule towards a desirable property, can be achieved by mutagenizing one or more parental molecular templates and by identifying any desirable molecules among the progeny molecules. Currently available technologies in directed evolution include methods for achieving stochastic (i.e. random) mutagenesis and methods for achieving non-stochastic (non-random) mutagenesis. However, critical shortfalls in both types of methods are identified in the instant disclosure.

In prelude, it is noteworthy that it may be argued philosophically by some that all mutagenesis – if considered from an objective point of view – is non-stochastic; and furthermore that the entire universe is undergoing a process that – if considered from an objective point of view – is non-stochastic. Whether this is true is outside of the scope of the instant consideration. Accordingly, as used herein, the terms “randomness”, “uncertainty”, and “unpredictability” have subjective meanings, and the knowledge, particularly the predictive knowledge, of the designer of an experimental process is a determinant of whether the process is stochastic or non-stochastic.

By way of illustration, stochastic or random mutagenesis is exemplified by a situation in which a progenitor molecular template is mutated (modified or changed) to yield a set of progeny molecules having mutation(s) that are not predetermined. Thus, in an in vitro stochastic mutagenesis reaction, for example, there is not a particular predetermined product whose production is intended; rather there is an uncertainty – hence randomness – regarding the exact nature of the mutations achieved, and thus also regarding the products generated. In contrast, non-stochastic or non-random mutagenesis is exemplified by a situation in which a progenitor molecular template is mutated (modified or changed) to yield a progeny molecule having one or more predetermined mutations. It is appreciated that the presence of background products in some quantity is a reality in many reactions where molecular processing occurs, and the presence of these background products does not detract from the non-stochastic nature of a mutagenesis process having a predetermined product.

Thus, as used herein, stochastic mutagenesis is manifested in processes such as error-prone PCR and stochastic shuffling, where the mutation(s) achieved are random or not predetermined. In contrast, as used herein, non-stochastic mutagenesis is manifested in instantly disclosed processes such as gene site-saturation mutagenesis and synthetic ligation reassembly, where the exact chemical structure(s) of the intended product(s) are predetermined.

In brief, existing mutagenesis methods that are non-stochastic have been serviceable in generating from one to only a very small number of predetermined mutations per method application, and thus produce per method application from one to only a few progeny molecules that have predetermined molecular structures. Moreover, the types of mutations currently available by the application of these non-stochastic methods are also limited, and thus so are the types of progeny mutant molecules.

In contrast, existing methods for mutagenesis that are stochastic in nature have been serviceable for generating somewhat larger numbers of mutations per method application — though in a random fashion & usually with a large but unavoidable contingency of undesirable background products. Thus, these existing stochastic methods can produce per method application larger numbers of progeny molecules, but that have undetermined molecular structures. The types of mutations that can be achieved by application of these current stochastic methods are also limited, and thus so are the types of progeny mutant molecules.

It is instantly appreciated that there is a need for the development of non-stochastic mutagenesis methods that:

- 1) Can be used to generate large numbers of progeny molecules that have predetermined molecular structures;
- 2) Can be used to readily generate more types of mutations;
- 3) Can produce a correspondingly larger variety of progeny mutant molecules;
- 4) Produce decreased unwanted background products;
- 5) Can be used in a manner that is exhaustive of all possibilities; and
- 6) Can produce progeny molecules in a systematic & non-repetitive way.

The instant invention satisfies all of these needs.

Directed Evolution Supplements Natural Evolution: Natural evolution has been a springboard for directed or experimental evolution, serving both as a reservoir of methods to be mimicked and of molecular templates to be mutagenized. It is appreciated that, despite its intrinsic process-related limitations (in the types of favored &/or allowed mutagenesis processes) and in its speed, natural evolution has had the advantage of having been in process for millions of years & and throughout a wide diversity of environments. Accordingly, natural evolution (molecular mutagenesis and selection in nature) has resulted

in the generation of a wealth of biological compounds that have shown usefulness in certain commercial applications.

However, it is instantly appreciated that many unmet commercial needs are discordant with any evolutionary pressure &/or direction that can be found in nature. Moreover, it is often the case that when commercially useful mutations would otherwise be favored at the molecular level in nature, natural evolution often overrides the positive selection of such mutations, e.g. when there is a concurrent detriment to an organism as a whole (such as when a favorable mutation is accompanied by a detrimental mutation). Additionally, natural evolution is often slow, and favors fidelity in many types of replication. Additionally still, natural evolution often favors a path paved mainly by consecutive beneficial mutations while tending to avoid a plurality of successive negative mutations, even though such negative mutations may prove beneficial when combined, or may lead - through a circuitous route - to final state that is beneficial.

Moreover, natural evolution advances through specific steps (e.g. specific mutagenesis and selection processes), with avoidance of less favored steps. For example, many nucleic acids do not reach close enough proximity to each other in a operative environment to undergo chimerization or incorporation or other types of transfers from one species to another. Thus, e.g., when sexual intercourse between 2 particular species is avoided in nature, the chimerization of nucleic acids from these 2 species is likewise unlikely, with parasites common to the two species serving as an example of a very slow passageway for inter-molecular encounters and exchanges of DNA. For another example, the generation of a molecule causing self-toxicity or self-lethality or sexual sterility is avoided in nature. For yet another example, the propagation of a molecule having no particular immediate benefit to an organism is prone to vanish in subsequent generations of the organism. Furthermore, e.g., there is no selection pressure for improving the performance of molecule under conditions other than those to which it is exposed in its endogenous environment; e.g. a cytoplasmic molecule is not likely to acquire functional features extending beyond what is required of it in the cytoplasm. Furthermore still, the propagation of a biological molecule is susceptible to any global detrimental effects - whether caused by itself or not - on its ecosystem. These and other characteristics greatly limit the types of mutations that can be propagated in nature.

On the other hand, directed (or experimental) evolution - particularly as provided herein - can be performed much more rapidly and can be directed in a more streamlined

manner at evolving a predetermined molecular property that is commercially desirable where nature does not provide one &/or is not likely to provide. Moreover, the directed evolution invention provided herein can provide more wide-ranging possibilities in the types of steps that can be used in mutagenesis and selection processes. Accordingly, using templates harvested from nature, the instant directed evolution invention provides more wide-ranging possibilities in the types of progeny molecules that can be generated and in the speed at which they can be generated than often nature itself might be expected to in the same length of time.

In a particular exemplification, the instantly disclosed directed evolution methods can be applied iteratively to produce a lineage of progeny molecules (e.g. comprising successive sets of progeny molecules) that would not likely be propagated (i.e., generated &/or selected for) in nature, but that could lead to the generation of a desirable downstream mutagenesis product that is not achievable by natural evolution.

Previous Directed Evolution Methods Are Suboptimal:

Mutagenesis has been attempted in the past on many occasions, but by methods that are inadequate for the purpose of this invention. For example, previously described non-stochastic methods have been serviceable in the generation of only very small sets of progeny molecules (comprised often of merely a solitary progeny molecule). By way of illustration, a chimeric gene has been made by joining 2 polynucleotide fragments using compatible sticky ends generated by restriction enzyme(s), where each fragment is derived from a separate progenitor (or parental) molecule. Another example might be the mutagenesis of a single codon position (i.e. to achieve a codon substitution, addition, or deletion) in a parental polynucleotide to generate a single progeny polynucleotide encoding for a single site-mutagenized polypeptide.

Previous non-stochastic approaches have only been serviceable in the generation of but one to a few mutations per method application. Thus, these previously described non-stochastic methods thus fail to address one of the central goals of this invention, namely the exhaustive and non-stochastic chimerization of nucleic acids. Accordingly previous non-stochastic methods leave untapped the vast majority of the possible point mutations, chimerizations, and combinations thereof, which may lead to the generation of highly desirable progeny molecules.

In contrast, stochastic methods have been used to achieve larger numbers of point mutations and/or chimerizations than non-stochastic methods; for this reason, stochastic

methods have comprised the predominant approach for generating a set of progeny molecules that can be subjected to screening, and amongst which a desirable molecular species might hopefully be found. However, a major drawback of these approaches is that – because of their stochastic nature – there is a randomness to the exact components in each set of progeny molecules that is produced. Accordingly, the experimentalist typically has little or no idea what exact progeny molecular species are represented in a particular reaction vessel prior to their generation. Thus, when a stochastic procedure is repeated (e.g. in a continuation of a search for a desirable progeny molecule), the re-generation and re-screening of previously discarded undesirable molecular species becomes a labor-intensive obstruction to progress, causing a circuitous – if not circular – path to be taken. The drawbacks of such a highly suboptimal path can be addressed by subjecting a stochastically generated set of progeny molecules to a labor-incurring process, such as sequencing, in order to identify their molecular structures, but even this is an incomplete remedy.

Moreover, current stochastic approaches are highly unsuitable for comprehensively or exhaustively generating all the molecular species within a particular grouping of mutations, for attributing functionality to specific structural groups in a template molecule (e.g. a specific single amino acid position or a sequence comprised of two or more amino acids positions), and for categorizing and comparing specific grouping of mutations. Accordingly, current stochastic approaches do not inherently enable the systematic elimination of unwanted mutagenesis results, and are, in sum, burdened by too many inherently shortcomings to be optimal for directed evolution.

In a non-limiting aspect, the instant invention addresses these problems by providing non-stochastic means for comprehensively and exhaustively generating all possible point mutations in a parental template. In another non-limiting aspect, the instant invention further provides means for exhaustively generating all possible chimerizations within a group of chimerizations. Thus, the aforementioned problems are solved by the instant invention.

Specific shortfalls in the technological landscape addressed by this invention include:

- 1) Site-directed mutagenesis technologies, such as sloppy or low-fidelity PCR, are ineffective for systematically achieving at each position (site) along a polypeptide sequence the full (saturated) range of possible mutations (i.e. all possible amino acid substitutions).
- 2) There is no relatively easy systematic means for rapidly analyzing the large amount of information that can be contained in a molecular sequence and in the potentially

colossal number of progeny molecules that could be conceivably obtained by the directed evolution of one or more molecular templates.

3) There is no relatively easy systematic means for providing comprehensive empirical information relating structure to function for molecular positions.

4) There is no easy systematic means for incorporating internal controls, such as positive controls, for key steps in certain mutagenesis (e.g. chimerization) procedures.

5) There is no easy systematic means to select for a specific group of progeny molecules, such as full-length chimeras, from among smaller partial sequences.

An exceedingly large number of possibilities exist for the purposeful and random combination of amino acids within a protein to produce useful hybrid proteins and their corresponding biological molecules encoding for these hybrid proteins, i.e., DNA, RNA. Accordingly, there is a need to produce and screen a wide variety of such hybrid proteins for a desirable utility, particularly widely varying random proteins.

The complexity of an active sequence of a biological macromolecule (e.g., polynucleotides, polypeptides, and molecules that are comprised of both polynucleotide and polypeptide sequences) has been called its information content ("IC"), which has been defined as the resistance of the active protein to amino acid sequence variation (calculated from the minimum number of invariable amino acids (bits) required to describe a family of related sequences with the same function). Proteins that are more sensitive to random mutagenesis have a high information content.

Molecular biology developments, such as molecular libraries, have allowed the identification of quite a large number of variable bases, and even provide ways to select functional sequences from random libraries. In such libraries, most residues can be varied (although typically not all at the same time) depending on compensating changes in the context. Thus, while a 100 amino acid protein can contain only 2,000 different mutations, 20^{100} sequence combinations are possible.

Information density is the IC per unit length of a sequence. Active sites of enzymes tend to have a high information density. By contrast, flexible linkers of information in enzymes have a low information density.

Current methods in widespread use for creating alternative proteins in a library format are error-prone polymerase chain reactions and cassette mutagenesis, in which the specific region to be optimized is replaced with a synthetically mutagenized

oligonucleotide. In both cases, a substantial number of mutant sites are generated around certain sites in the original sequence.

Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. In a mixture of fragments of unknown sequence, error-prone PCR can be used to mutagenize the mixture. The published error-prone PCR protocols suffer from a low processivity of the polymerase. Therefore, the protocol is unable to result in the random mutagenesis of an average-sized gene. This inability limits the practical application of error-prone PCR. Some computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the large-scale block changes that are required for continued and dramatic sequence evolution. Further, the published error-prone PCR protocols do not allow for amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical application. In addition, repeated cycles of error-prone PCR can lead to an accumulation of neutral mutations with undesired results, such as affecting a protein's immunogenicity but not its binding affinity.

In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping them into families, arbitrarily choosing a single family, and reducing it to a consensus motif. Such motif is re-synthesized and reinserted into a single gene followed by additional selection. This step process constitutes a statistical bottleneck, is labor intensive, and is not practical for many rounds of mutagenesis.

Error-prone PCR and oligonucleotide-directed mutagenesis are thus useful for single cycles of sequence fine-tuning, but rapidly become too limiting when they are applied for multiple cycles.

Another limitation of error-prone PCR is that the rate of down-mutations grows with the information content of the sequence. As the information content, library size, and mutagenesis rate increase, the balance of down-mutations to up-mutations will statistically prevent the selection of further improvements (statistical ceiling).

In cassette mutagenesis, a sequence block of a single template is typically replaced by a (partially) randomized sequence. Therefore, the maximum information content that can be obtained is statistically limited by the number of random sequences (i.e., library size). This eliminates other sequence families which are not currently best, but which may have greater long term potential.

Also, mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round. Thus, such an approach is tedious and impractical for many rounds of mutagenesis.

Thus, error-prone PCR and cassette mutagenesis are best suited, and have been widely used, for fine-tuning areas of comparatively low information content. One apparent exception is the selection of an RNA ligase ribozyme from a random library using many rounds of amplification by error-prone PCR and selection.

In nature, the evolution of most organisms occurs by natural selection and sexual reproduction. Sexual reproduction ensures mixing and combining of the genes in the offspring of the selected individuals. During meiosis, homologous chromosomes from the parents line up with one another and cross-over part way along their length, thus randomly swapping genetic material. Such swapping or shuffling of the DNA allows organisms to evolve more rapidly.

In recombination, because the inserted sequences were of proven utility in a homologous environment, the inserted sequences are likely to still have substantial information content once they are inserted into the new sequence.

Theoretically there are 2,000 different single mutants of a 100 amino acid protein. However, a protein of 100 amino acids has 20^{100} possible sequence combinations, a number which is too large to exhaustively explore by conventional methods. It would be advantageous to develop a system which would allow generation and screening of all of these possible combination mutations.

Some workers in the art have utilized an *in vivo* site specific recombination system to generate hybrids of combine light chain antibody genes with heavy chain antibody genes for expression in a phage system. However, their system relies on specific sites of recombination and is limited accordingly. Simultaneous mutagenesis of antibody CDR regions in single chain antibodies (scFv) by overlapping extension and PCR have been reported.

Others have described a method for generating a large population of multiple hybrids using random *in vivo* recombination. This method requires the recombination of two different libraries of plasmids, each library having a different selectable marker. The method is limited to a finite number of recombinations equal to the number of selectable markers existing, and produces a concomitant linear increase in the number of marker genes linked to the selected sequence(s).

In vivo recombination between two homologous, but truncated, insect-toxin genes on a plasmid has been reported as a method of producing a hybrid gene. The *in vivo* recombination of substantially mismatched DNA sequences in a host cell having defective mismatch repair enzymes, resulting in hybrid molecule formation has been reported.

C - SUMMARY OF THE INVENTION

This invention relates generally to the field of cellular and whole organism engineering. Specifically, this invention relates to a cellular transformation, directed evolution, and screening method for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable.

In one embodiment, this invention is directed to a method of producing an improved organism having a desirable trait to by: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence of said improved organism. This invention provides that any of steps a), b), and c) can be further repeated in any particular order and any number of times; accordingly, this invention specifically provides methods comprised of any iterative combination of steps a), b), and c), with a number of iterations.

In another embodiment, this invention is directed to a method of producing an improved organism having a desirable trait to by: a) obtaining an initial population of organisms, which can be a clonal population or otherwise, b) generating a set of mutagenized organisms each having at least one genetic mutation, such that when all the

genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations c) detecting the manifestation of at least two genetic mutations, and d) introducing at least two detected genetic mutations into one organism. Additionally, this invention provides that any of steps a), b), c), and d) can be further repeated in any particular order and any number of times; accordingly, this invention specifically provides methods comprised of any iterative combination of steps a), b), c), and d), with a total number of iterations can be from one up to one million, including specifically every integer value in between.

In a preferred aspect of embodiments specified herein the step of b) generating a second set of mutagenized organisms is comprised of generating a plurality of organisms, each of which organisms has a particular transgenic mutation.

As used herein, **"generating a set of mutagenized organisms having genetic mutations"** can be achieved by any means known in the art to mutagenized including any radiation known to mutagenized, such as ionizing and ultra violet. Further examples of serviceable mutagenizing methods include site-saturation mutagenesis, transposon-based methods, and homologous recombination.

"Combining" means incorporating a plurality of different genetic mutations in the genetic makeup (e.g. the genome) of the same organism; and methods to achieve this "combining" step including sexual recombination, homologous recombination, and transposon-based methods.

As used herein, an **"initial population of organisms"** means a **"working population of organisms"**, which refers simply to a population of organisms with which one is working, and which is comprised of at least one organism. An **"initial population of organisms"** which can be a clonal population or otherwise.

Accordingly, in step 1) an **"initial population of organisms"** may be a population of multicellular organisms or of unicellular organisms or of both. An **"initial population of organisms"** may be comprised of unicellular organisms or multicellular organisms or both. An **"initial population of organisms"** may be comprised of prokaryotic organisms or

eukaryotic organisms or both. This invention provides that an "initial population of organisms" is comprised of at least one organism, and preferred embodiments include at least that .

By "organism" is meant any biological form or thing that is capable of self replication or replication in a host. Examples of "organisms" include the following kinds of organisms (which kinds are not necessarily mutually-exclusive): animals, plants, insects, cyanobacteria, microorganisms, fungi, bacteria, eukaryotes, prokaryotes, mycoplasma, viral organisms (including DNA viruses, RNA viruses), and prions.

Non-limiting particularly preferred examples of kinds of "organisms" also include Archaea (archaeobacteria) and Bacteria (eubacteria). Non-limiting examples of Archaea (archaeobacteria) include Crenarchaeota, Euryarchaeota, and Korarchaeota. Non-limiting examples Bacteria (eubacteria) include Aquificales, CFB/Green sulfur bacteria group, Chlamydiales/Verrucomicrobia group, Chrysiogenes group, Coprothermobacter group, Cyanobacteria & chloroplasts, Cytophaga/Flexibacter /Bacteriodes group, Dictyoglomus group, Fibrobacter/Acidobacteria group, Firmicutes, Flexistipes group, Fusobacteria, Green non-sulfur bacteria, Nitrospira group, Planctomycetales, Proteobacteria, Spirochaetales, Synergistes group, Thermodesulfobacterium group, Thermotogales, Thermus/Deinococcus group. As non-limiting examples, particularly preferred kinds of organisms include Aquifex, Aspergillus, Bacillus, Clostridium, E. coli, Lactobacillus, Mycobacterium, Pseudomonas, Streptomyces, and Thermotoga. As additional non-limiting examples, particularly preferred organisms include cultivated organisms such as CHO, VERO, BHK, HeLa, COS, MDCK, Jurkat, HEK-293, and WI38. Particularly preferred non-limiting examples of organisms further include host organisms that are serviceable for the expression of recombinant molecules. Organisms further include primary cultures (e.g. cells from harvested mammalian tissues), immortalized cells, all cultivated and culturable cells and multicellular organisms, and all uncultivated and unculturable cells and multicellular organisms.

In a preferred embodiment, knowledge of genomic information is useful for performing the claimed methods; thus, this invention provides the following as preferred but non-limiting examples of organisms that are particularly serviceable for this invention,

because there is a significant amount of - if not complete - genomic sequence information (in terms of primary sequence &/or annotation) for these organisms: Human, Insect (e.g. *Drosophila melanogaster*), Higher plants (e.g. *Arabidopsis thaliana*), Protozoan (e.g. *Plasmodium falciparum*), Nematode (e.g. *Caenorhabditis elegans*), Fungi (e.g. *Saccharomyces cerevisiae*), Proteobacteria gamma subdivision (e.g. *Escherichia coli* K-12, *Haemophilus influenzae* Rd, *Xylella fastidiosa* 9a5c, *Vibrio cholerae* El Tor N16961, *Pseudomonas aeruginosa* PA01, *Buchnera* sp. APS), Proteobacteria beta subdivision (e.g. *Neisseria meningitidis* MC58 (serogroup B), *Neisseria meningitidis* Z2491 (serogroup A)), Proteobacteria other subdivisions (e.g. *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Campylobacter jejuni* NCTC11168, *Rickettsia prowazekii*), Gram-positive bacteria (e.g. *Bacillus subtilis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Ureaplasma urealyticum*, *Mycobacterium tuberculosis* H37Rv), Chlamydia (e.g. *Chlamydia trachomatis* serovar D, *Chlamydia muridarum* (*Chlamydia trachomatis* MoPn), *Chlamydia pneumoniae* CWL029, *Chlamydia pneumoniae* AR39, *Chlamydia pneumoniae* J138), Spirochete (e.g. *Borrelia burgdorferi* B31, *Treponema pallidum*), Cyanobacteria (e.g. *Synechocystis* sp. PCC6803), Radioresistant bacteria (e.g. *Deinococcus radiodurans* R1), Hyperthermophilic bacteria (e.g. *Aquifex aeolicus* VF5, *Thermotoga maritima* MSB8), and Archaea (e.g. *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum* deltaH, *Archaeoglobus fulgidus*, *Pyrococcus horikoshii* OT3, *Pyrococcus abyssi*, *Aeropyrum pernix* K1).

Non-limiting particularly preferred examples of kinds of plant "organisms" include those listed in Table 1.

Table 1. Non-limiting examples of plant organisms and sources of transgenic molecules (e.g. nucleic acids & nucleic acid products)

1. Alfalfa	39. Pepper
2. Amelanchier laevis	40. Persimmon
3. Apple	41. Petunia
4. Arab. thaliana	42. Pine
5. Arabidopsis	43. Pineapple
6. Aspergillus flavus	44. Pink bollworm
7. Barley	45. Plum
8. Beet	46. Poplar
9. Belladonna	47. Potato
10. Brassica oleracea	48. Pseudomonas

11. Carrot	49. <i>Pseudomonas putida</i>
12. Chrysanthemum	50. <i>Pseudomonas syringae</i>
13. Cichorium intybus	51. Rapeseed
14. Clavibacter	52. Rhizobium
15. Clavibacter xyli	53. Rhizobium etli
16. Coffee	54. Rhizobium fredii
17. Corn	55. Rhizobium leguminosarum
18. Cotton	56. Rhizobium meliloti
19. Cranberry	57. Rice
20. Creeping bentgrass	58. Rubus idaeus
21. Cryphonectria parasitica	59. Spruce
22. Eggplant	60. Soybean
23. Festuca arundinacea	61. Squash
24. Fusarium graminearum	62. Squash-cucumber
25. Fusarium moniliforme	63. Squash-cucurbita texana
26. Fusarium sporotrichioides	64. Strawberry
27. Gladiolus	65. Sugarcane
28. Grape	66. Sunflower
29. Heterorhabditis bacteriophora	67. Sweet potato
30. Kentucky bluegrass	68. Sweetgum
31. Lettuce	69. TMV
32. Melon	70. Tobacco
33. Oat	71. Tomato
34. Onion	72. Walnut
35. Papaya	73. Watermelon
36. Pea	74. Wheat
37. Peanut	75. Xanthomonas
38. Pelargonium	76. Xanthomonas campestris

As used herein, the meaning of "generating a set of mutagenized organisms having genetic mutations" includes the steps of substituting, deleting, as well as introducing a nucleotide sequence into organism; and this invention provides a nucleotide sequence that serviceable for this purpose may be a single-stranded or double-stranded and the fact that its length may be from one nucleotide up to 10,000,000,000 nucleotides in length including specifically every integer value in between.

A mutation in an organism includes any alteration in the structure of one or more molecules that encode the organism. These molecules include nucleic acid, DNA, RNA, prionic molecules, and may be exemplified by a variety of molecules in an organism such as a DNA that is genomic, episomal, or nucleic, or by a nucleic acid that is vectoral (e.g. viral, cosmid, phage, phagemid).

In one aspect, as used herein, a "set of substantial genetic mutations" is preferably a disruption (e.g. a functional knock-out) of at least about 15 to about 150,000

genomic locations or nucleotide sequences (e.g. genes, promoters, regulatory sequences, codons etc.), including specifically every integer value in between. In another aspect, as used herein, a **“set of substantial genetic mutations”** is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 15 to about 150,000 genes, including specifically every integer value in between. Corresponding to another aspect, as used herein, a **“set of substantial genetic mutations”** is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 15 to about 150,000 gene products &/or phenotypes &/or traits, including specifically every integer value in between.

In another aspect, as used herein, a **“set of substantial genetic mutations”** with respect to an organism (or type of organism) is preferably a disruption (e.g. a functional knock-out) of at least about 1% to about 100% of genomic locations or nucleotide sequences (e.g. genes, promoters, regulatory sequences, codons etc.) in the organism (or type of organism), including specifically percentages of every integer value in between. In another aspect, as used herein, a **“set of substantial genetic mutations”** is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 1% to about 100% of genes in an organism (or type of organism), including specifically percentages of every integer value in between. Corresponding to another aspect, as used herein, a **“set of substantial genetic mutations”** is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 1% to about 100% of the gene products &/or phenotypes &/or traits of an organism (or type of organism), including specifically every integer value in between.

In yet another aspect, as used herein, a **“set of substantial genetic mutations”** is preferably an introduction or deletion of at least about 15 to 150,000 genes promoters or other nucleotide sequences (where each sequence is from 1 base to 10,000,000 bases), including specifically every integer value in between. For example, one can introduce a library of at least about 15 to 150,000 nucleotides (genes or promoters) produced by “site-

saturation mutagenesis" &/or by "ligation reassembly" (including any specific aspect thereof provided herein) into an "initial population of organisms".

It is provided that wherever the manipulation of a plurality of "genes" is mentioned herein, gene pathways (e.g. that ultimately lead to the production of small molecules) are also included. It is appreciated herein that knocking-out, altering expression level, and altering expression pattern can be achieved, by non-limiting exemplification, by mutagenizing a nucleotide sequence corresponding gene as well as a corresponding promoter that affects the expression of the gene.

As used herein, a "mutagenized organism" includes any organism that has been altered by a genetic mutation.

A "genetic mutation" can be, by way of non-limiting and non-mutually exclusive exemplification, and change in the nucleotide sequence (DNA or RNA) with respect to genomic, extra-genomic, episomal, mitochondrial, and any nucleotide sequence associated with (e.g. contained within or considered part of) an organism..

According to this invention, detecting the manifestation of a "genetic mutation" means "detecting the manifestation of a detectable parameter", including but not limited to a change in the genomic sequence. Accordingly, this invention provides that a step of sequencing (&/or annotating) of and organism's genomic DNA is necessary for some methods of this invention, and exemplary but non-limiting aspects of this sequencing (&/or annotating) step are provided herein.

A detectable "trait", as used herein, is any detectable parameter associated with the organism. Accordingly, such a detectable "parameter" includes, by way of non-limiting exemplification, any detectable "nucleotide knock-in", any detectable "nucleotide knock-outs", any detectable "phenotype", and any detectable "genotype". By way of further illustration, a "trait" includes any substance produced or not produced by the organism. Accordingly, a "trait" includes viability or non-viability, behavior, growth rate, size, morphology. "Trait" includes increased (or alternatively decreased) expression of a gene product or gene pathway product. "Trait" also includes small molecule production

(including vitamins, antibiotics), herbicide resistance, drought resistance, pest resistance, production of any recombinant biomolecule (ie.g. vaccines, enzymes, protein therapeutics, chiral enzymes). Additional examples of serviceable traits for this invention are shown in Table 2.

TABLE 2 – Non-limiting examples of serviceable genes, gene products, phenotypes, or traits according to the methods of this invention (e.g. knockouts, knockins, increased or decreased expression level, increased or decreased expression pattern)

Table 2 - Part 1. Non-limiting examples of genes or gene products

1. 17 kDa protein	53. Cecropin
2. 3-hydroxy-3-methylglutaryl CoenzymeA reductase	54. Cecropin B
3. 4-Coumarate:CoA ligase knockout	55. Cellulose binding protein
4. 60 kDa protein	56. Chalcone synthase knockout
5. Ac transposable element	57. Chitinase
6. ACC deaminase	58. Chitobiosidase
7. ACC oxidase knockout	59. Chloramphenicol acetyltransferase
8. ACC synthase	60. Cholera toxin B
9. ACC synthase knockout	61. Choline oxidase
10. Acetohydroxyacid synthase variant	62. Cinnamate 4-hydroxylase
11. Acetolactate synthase	63. Cinnamate 4-hydroxylase knockout
12. Acetyl CoA carboxylase	64. Coat protein
13. ACP acyl-ACP thioesterase	65. Coat protein knockout
14. ACP thioesterase	66. Conglycinin
15. Acyl CoA reductase	67. CryIA
16. Acyl-ACP knockout	68. CryIAb
17. Acyl-ACP desaturase	69. CryIAc
18. Acyl-ACP desaturase knockout	70. CryIB
19. Acyl-ACP thioesterase	71. CryIIA
20. ADP glucose pyrophosphorylase	72. CryIIIA
21. ADP glucose pyrophosphorylase knockout	73. CryVIA
22. Agglutinin	74. Cyclin dependent kinase
23. Aleurone 1	75. Cyclodextrin glycosyltransferase
24. Alpha hordothionin	76. Cylindrical inclusion protein
25. Alpha-amylase	77. Cystathionine synthase
26. Alpha-hemoglobin	78. Delta-12 desaturase
27. Aminoglycoside 3'-adenylyltransferase	79. Delta-12 desaturase knockout
28. Amylase	80. Delta-12 saturase
29. Anionic peroxidase	81. Delta-12 saturase knockout
30. Antibody	82. Delta-15 desaturase
31. Antifungal protein	83. Delta-15 desaturase knockout
32. Antithrombin	84. Delta-9 desaturase
33. Antitrypsin	85. Delta-9 desaturase knockout
34. Antiviral protein	86. Deoxyhypusine synthase (DHS)
35. Aspartokinase	87. Deoxyhypusine synthase knockout
36. Attacin E	88. Diacylglycerol acetyl transferase
37. B1 regulatory gene	89. Dihydrodipicolinate synthase
38. B-1,3-glucanase knockout	90. Dihydrofolate reductase
39. B-1,4-endoglucanase knockout	91. Diphtheria toxin A
40. Bacteropsin	92. Disease resistance response gene 49
41. Barnase	93. Double stranded ribonuclease
42. Barstar	94. Ds transposable element
43. Beta-hemoglobin	95. Elongase

44. B-glucuronidase	96. EPSPS
45. C1 knockout	97. Ethylene forming enzyme knockout
46. C1 regulatory gene	98. Ethylene receptor protein
47. C2 knockout	99. Ethylene receptor protein knockout
48. C3 knockout	100. Fatty acid elongase
49. Caffeate O-methyltransferase	101. Fluorescent protein
50. Caffeate O-methyltransferase knockout	102. G glycoprotein
51. Caffeoyl CoA O-methyltransferase knockout	103. Galactanase
52. Casein	104. Galanthus nivalis agglutinin

Table 2 – Part I. (continued) Non-limiting examples of transgenic genes & gene knockouts

105. Genome-linked protein	157. Omega 3 desaturase knockout
106. Glucanase	158. Omega 6 desaturase
107. Glucanase knockout	159. Omega 6 desaturase knockout
108. Glucose oxidase	160. O-methyltransferase
109. Glutamate dehydrogenase	161. Osmotin
110. Glutamine binding protein	162. Oxalate oxidase
111. Glutamine synthetase	163. Par locus
112. Glutenin	164. Pathogenesis protein 1a
113. Glycerol-3-phosphate acetyl transferase	165. Pectate lyase
114. Glyphosate oxidoreductase	166. Pectin esterase
115. Glyphosate oxidoreductase	167. Pectin esterase knockout
116. Green fluorescent protein	168. Pectin methylesterase
117. Helper component	169. Pectin methylesterase knockout
118. Hemicellulase	170. Pentenlypyrophosphate isomerase
119. Hup locus	171. Phosphinothricin
120. Hygromycin phosphotransferase	172. Phosphinothricin acetyl transferase
121. Hyoscamine 6B-hydroxylase	173. Phytochrome A
122. IAA monooxygenase	174. Phytoene synthase
123. Invertase	175. Phleomycin binding protein
124. Invertase knockout	176. Polygalacturonase
125. Isopentenyl transferase	177. Polygalacturonase knockout
126. Ketoacyl-ACP synthase	178. Polygalacturonase inhibitor protein
127. Ketoacyl-ACP synthase knockout	179. Prf regulatory gene
128. Larval serum protein	180. Prosystemin
129. Leafy homeotic regulatory gene	181. Protease
130. Lectin	182. Protein A
131. Lignin peroxidase	183. Protein kinase
132. Luciferase	184. Proteinase inhibitor I
133. Lysine-2 gene	185. Pti5 transcription factor
134. Lysophosphatidic acid acetyl transferase	186. R regulatory gene
135. Lysozyme	187. Receptor kinase
136. Mabinlin	188. Recombinase
137. Male sterility protein	189. Reductase
138. Metallothionein	190. Replicase
139. Modified ethylene receptor protein	191. Resveratrol synthase
140. Modified ethylene receptor protein knockout	192. Ribonuclease
141. Monooxygenase	193. rolc
142. Movement protein	194. Rol hormone gene
143. Movement protein nonfunctional	195. S-adenosylmethione decarboxylase
144. N gene for TMV resistance	196. S-adenosylmethione hydrolase
145. N-acetyl glucosidase	197. S-adenosylmethionine transferase
146. Nitrilase	198. Salicylate hydroxylase
147. Nopaline synthase	199. Satellite RNA
148. Notch	200. Seed storage protein
149. NptII	201. Serine-threonine protein kinase
150. Nuclear inclusion protein a	202. Serum albumin
151. Nuclear inclusion protein b	203. Shrunk 2
152. Nucleocapsid	204. Sorbitol dehydrogenase
153. Nucleoprotein	205. Sorbitol synthase

154. O-acyl transferase	206. Stilbene synthase
155. Oleoyl-ACP thioesterase	207. Storage protein
156. Omega 3 desaturase	208. Sucrose phosphate synthase

Table 2 – Part 1.(continued) Non-limiting examples of transgenic genes & gene knockouts

209. Systemic acquired resistance gene 8.2	219. Trichosanthin
210. Tetracycline binding protein	220. Trifolixotoxin
211. Thioesterase (x2)	221. Trypsin inhibitor
212. Thiolase	222. T-URF13 mitochondrial
213. TobRB7	223. UDP glucose glucosyltransferase
214. Transcriptional activator	224. Violaxanthin de-epoxidase
215. Transposon Tn5	225. Violaxanthin de-epoxidase knockout
216. Trehalase	226. Wheat germ agglutinin
217. Trehalase knockout	227. Xanthosine-N7-methyltransferase knockout
218. Trichodiene synthase	228. Zein storage protein

Table 2 – Part 2. Non-limiting examples of input traits/phenotypes

1. 2,4-D tolerant	52. Flowering time altered
2. Alternaria resistant	53. Frogeye leaf spot resistant
3. Altered amino acid composition	54. Fruit ripening altered
4. Alternaria solani resistant	55. Fruit ripening delayed
5. Ammonium assimilation increased	56. Fruit rot resistant
6. AMV resistant	57. Fruit solids increased
7. Aphid resistant	58. Fruit sweetness increased
8. Apple scab resistant	59. Fungal post-harvest resistant
9. Aspergillus resistant	60. Fungal resistant
10. B-1,4-endoglucanase	61. Fungal resistant general
11. Bacterial leaf blight resistant	62. Fusarium resistant
12. Bacterial speck resistant	63. Glyphosate tolerant
13. BCTV resistant	64. Growth rate altered
14. Blackspot bruise resistant	65. Growth rate reduced
15. BLRV resistant	66. Heat stable glucanase produced
16. BNYYV Resistant	67. Hordothionin produced
17. Botrytis cinerea resistant	68. Imidazolinone tolerant
18. Botrytis resistant	69. Insect resistant general
19. BPMV resistant	70. Kanamycin resistant
20. Bromoxynil tolerant	71. Lepidopteran resistant
21. BYDV resistant	72. Lesser cornstalk borer resistant
22. BYMV resistant	73. LMV resistant
23. Carbohydrate metabolism altered	74. Loss of systemic resistance
24. Cell wall altered	75. Male sterile
25. Chlorsulfuron tolerant	76. Marssonina resistant
26. Clavibacter resistant	77. MCDV resistant
27. CLRV resistant	78. MCMV resistant
28. CMV resistant	79. MDMV resistant
29. Cold tolerant	80. MDMV-B resistant
30. Coleopteran resistant	81. Mealybug wilt virus resistant
31. Colletotrichum resistant	82. Melampsora resistant
32. Colorado potato beetle resistant	83. Melodgyne resistant
33. Constitutive expression of glutamine synthetase	84. Methotrexate resistant
34. Corynebacterium sepedonicum resistant	85. Mexican Rice Borer resistant
35. Cottonwood leaf beetle resistant	86. Nucleocapsid protein produced
36. Crown gall resistant	87. Oblique banded leafroller resistant
37. Crown rot resistant	88. PEMV resistant
38. Cucumovirus resistant	89. PeSV resistant
39. Cutting rootability increased	90. Phoma resistant
40. Downy mildew resistant	91. Phosphinothricin tolerant
41. Drought tolerant	92. Phratara leaf beetle resistant

42. <i>Erwinia carotovora</i> resistant	93. <i>Phytophthora</i> resistant
43. Ethylene production reduced	94. PLRV resistant
44. European Corn Borer resistant	95. Polyamine metabolism altered
45. Female sterile	96. Potyvirus resistant
46. Fenthion susceptible	97. Powdery mildew resistant
47. Fertility altered	98. PPV resistant
48. Fire blight resistant	99. <i>Pratylenchus vulnus</i> resistant
49. Flower and fruit abscission reduced	100. Proteinase inhibitors level constitutive
50. Flower and fruit set altered	101. PRSV resistant
51. Flowering altered	102. PRV resistant

Table 2 – Part 2.(continued)Non-limiting examples of transgenic input traits/phenotypes

103. PSbMV resistant	128. <i>Streptomyces scabies</i> resistant
104. <i>Pseudomonas syringae</i> resistant	129. Sulfonyleurea tolerant
105. PSTV resistant	130. Tetracycline binding protein produced
106. PVX resistant	131. TEV resistant
107. PVY resistant	132. <i>Thelaviopsis</i> resistant
108. RBDV resistant	133. TMV resistant
109. <i>Rhizoctonia</i> resistant	134. Tobamovirus resistant
110. <i>Rhizoctonia solani</i> resistant	135. ToMoV resistant
111. Ring rot resistance	136. ToMV resistant
112. Root-knot nematode resistant	137. Transposon activator
113. SbMV resistant	138. Transposon inserted
114. Sclerotinia resistant	139. TRV resistant
115. SCMV resistant	140. TSWV resistant
116. SCYL V resistant	141. TVMV resistant
117. Secondary metabolite increased	142. TYLCV resistant
118. Seed set reduced	143. Tyrosine level increased
119. Selectable marker	144. <i>Venturia</i> resistant
120. Senescence altered	145. <i>Verticillium dahliae</i> resistant
121. <i>Septoria</i> resistant	146. <i>Verticillium</i> resistant
122. Shorter stems	147. Visual marker
123. Soft rot fungal resistant	148. WMV2 resistant
124. Soft rot resistant	149. WSMV resistant
125. SqMV resistant	150. Yield increased
126. SrMV resistant	151. ZYMV resistant
127. Storage protein altered	

Table 2 – Part 3. Non-limiting examples of output traits/phenotypes

1. ACC oxidase level decreased	36. Oil profile altered
2. Altered lignin biosynthesis	37. Pectin esterase level reduced
3. B-1,4-endoglucanase	38. Pharmaceutical proteins produced
4. <i>Botrytis</i> resistant	39. Phosphinothricin tolerant
5. Carbohydrate metabolism altered	40. Phytoene synthase activity increased
6. Carotenoid content altered	41. Pigment metabolism altered
7. Cell wall altered	42. Polygalacturonase level reduced
8. CMV resistant	43. Processing characteristics altered
9. Coleopteran resistant	44. Prolonged shelf life
10. Dry matter content increased	45. Protein altered
11. Ethylene production reduced	46. Protein quality altered
12. Ethylene synthesis reduced	47. PRSV resistant
13. Fatty acid metabolism altered	48. Root-knot nematode resistant
14. Fire blight resistant	49. Sclerotinia resistant
15. Flower and fruit abscission reduced	50. Seed composition altered
16. Flower and fruit set altered	51. Seed methionine storage increased
17. Flowering time altered	52. Seed set reduced
18. Fruit firmness increased	53. Seed storage protein
19. Fruit pectin esterase levels decreased	54. Senescence altered (e.g. Shelf life increased)
20. Fruit ripening altered	55. Shorter stems

21. Fruit ripening delayed	56. Solids increased
22. Fruit solids increased	57. SqMV resistant
23. Fruit sugar profile altered	58. Starch level increased
24. Fruit sweetness increased	59. Starch metabolism altered
25. Glucuronidase expressing	60. Starch reduced
26. Heat stable glucanase produced	61. Sterols increased
27. Heavy metals sequestered	62. Storage protein altered
28. Hordothionin produced	63. Sugar alcohol levels increased
29. Improved fruit quality	64. Tetracycline binding protein produced
30. Industrial enzyme produced	65. Tyrosine level increased
31. Lepidopteran resistant	66. Verticillium resistant
32. Lysine level increased	67. Visual marker
33. Mealybug wilt virus resistant	68. WMV2 resistant
34. Methionine level increased	69. Yield increased
35. Nucleocapsid protein produced	70. ZYMV resistant

Table 2 - Part 4. Non-limiting examples of traits/phenotypes with agronomic properties

1. ACC oxidase level decreased	53. Industrial enzyme produced
2. Altered amino acid composition	54. Lignin levels decreased
3. Altered lignin biosynthesis	55. Lipase expressed in seeds
4. Altered maturing	56. Lysine level increased
5. Altered plant development	57. Male sterile
6. Aluminum tolerant	58. Male sterile reversible
7. Ammonium assimilation increased	59. Methionine level increased
8. Anthocyanin produced in seed	60. Modified growth characteristics
9. B-1,4-endoglucanase	61. Mycotoxin degradation
10. Calmodulin level altered	62. Nitrogen metabolism altered
11. Carbohydrate metabolism altered	63. Nucleocapsid protein produced
12. Carotenoid content altered	64. Oil profile altered
13. Cell wall altered	65. Oil quality altered
14. Cold tolerant	66. Oxidative stress tolerant
15. Constitutive expression of glutamine synthetase	67. Pectin esterase level reduced
16. Cutting root ability increased	68. Pharmaceutical proteins produced
17. Development altered	69. Photosynthesis enhanced
18. Drought tolerant	70. Phytoene synthase activity increased
19. Dry matter content increased	71. Pigment metabolism altered
20. Environmental stress reduced	72. Polyamine metabolism altered
21. Ethylene metabolism altered	73. Polygalacturonase level reduced
22. Ethylene production reduced	74. Pratylenchus vulnus resistant
23. Ethylene synthesis reduced	75. Processing characteristics altered
24. Fatty acid metabolism altered	76. Prolonged shelf life
25. Female sterile	77. Protein altered
26. Fenthion susceptible	78. Protein lysine level increased
27. Fertility altered	79. Protein quality altered
28. Fiber quality altered	80. Proteinase inhibitors level constitutive
29. Flower and fruit abscission reduced	81. Salt tolerance increased
30. Flower and fruit set altered	82. Seed composition altered
31. Flowering altered	83. Seed methionine storage increased
32. Flower color altered	84. Seed set reduced
33. Flowering time altered	85. Selectable marker
34. Fruit firmness increased	86. Senescence altered
35. Fruit pectin esterase and levels decreased	87. Shorter stems
36. Fruit polygalacturonase level decreased	88. Solids increased
37. Fruit ripening altered	89. Starch level increased
38. Fruit ripening delayed	90. Starch metabolism altered
39. Fruit solids increased	91. Starch reduced
40. Fruit sugar profile altered	92. Sterols increased
41. Fruit sweetness increased	93. Storage protein altered
42. Glucuronidase expressing	94. Stress tolerant
43. Growth rate altered	95. Sugar alcohol levels increased

44. Growth rate increased	96. Tetracycline binding protein produced
45. Growth rate reduced	97. Thermostable protein produced
46. Heat stable glucanase produced	98. Transposon activator
47. Heat tolerant	99. Transposon inserted
48. Heavy metals sequestered	100. Tyrosine level increased
49. Hordothionin produced	101. Visual marker
50. Improved fruit quality	102. Vivipary increased
51. Increased phosphorus	103. Yield increased
52. Increased stalk strength	

Table 2 – Part 5. Non-limiting examples of traits/phenotypes with product quality properties

1. 2,4-D tolerant	45. Melanin produced in cotton fibers
2. ACC oxidase level decreased	46. Metabolism altered
3. Altered amino acid composition	47. Methionine level increased
4. Altered lignin biosynthesis	48. Mycotoxin degradation
5. Anthocyanin produced in seed	49. Mycotoxin production inhibited
6. Antioxidant enzyme increased	50. Nicotine levels reduced
7. Auxin metabolism and increased tuber solids	51. Nitrogen metabolism altered
8. B-1,4-endoglucanase	52. Novel protein produced
9. Blackspot bruise resistant	53. Nutritional quality altered
10. Brown spot resistant	54. Oil profile altered
11. Bruising reduced	55. Oil quality altered
12. Caffeine levels reduced	56. Pectin esterase level reduced
13. Carbohydrate metabolism altered	57. Photosynthesis enhanced
14. Carotenoid content altered	58. Phytoene synthase activity increased
15. Cell wall altered	59. Pigment metabolism altered
16. Cold tolerant	60. Polyamine metabolism altered
17. Delayed softening	61. Polygalacturonase level reduced
18. Disulfides reduced in endosperm	62. Processing characteristics altered
19. Dry matter content increased	63. Prolonged shelf life
20. Ear mold resistant	64. Protein altered
21. Ethylene production reduced	65. Protein lysine level increased
22. Ethylene synthesis reduced	66. Protein quality altered
23. Extended flower life	67. Proteinase inhibitors level constitutive
24. Fatty acid metabolism altered	68. Rust resistant
25. Fiber quality altered	69. Seed composition altered
26. Fiber strength altered	70. Seed methionine storage increased
27. Flavor enhancer	71. Seed number increased
28. Flower and fruit abscission reduced	72. Seed quality altered
29. Fruit firmness increased	73. Seed set reduced
30. Fruit invertase level decreased	74. Seed weight increased
31. Fruit polygalacturonase level decreased	75. Senescence altered
32. Fruit ripening altered	76. Solids increased
33. Fruit ripening delayed	77. Starch level increased
34. Fruit solids increased	78. Starch metabolism altered
35. Fruit sugar profile altered	79. Starch reduced
36. Fruit sweetness increased	80. Steroidal glycoalkaloids reduced
37. Glyphosate tolerant	81. Sterols increased
38. Heat stable glucanase produced	82. Storage protein altered
39. Improved fruit quality	83. Sugar alcohol levels increased
40. Increased phosphorus	84. Thermostable protein produced
41. Increased protein levels	85. Tryptophan level increased
42. Lignin levels decreased	86. Tuber solids increased
43. Lysine level increased	87. Yield increased
44. Male sterile	

Table 2 – Part 6. Non-limiting examples of traits/phenotypes with herbicide tolerance properties

1. 2,4-D tolerant	11. Sulfonyleurea tolerant
2. Chloroacetanilide tolerant	12. Northern corn leaf blight resistant

3. Fertility altered	13. Herbicide tolerant
4. Protein altered	14. Isoxazole tolerant
5. Lignin levels decreased	15. Chlorsulfuron tolerant
6. Methionine level increased	16. Glyphosate tolerant
7. Bromoxynil tolerant	17. Lepidopteran resistant
8. Metabolism altered	18. Phosphinothricin tolerant
9. Imidazole tolerant	19. Sulfonylurea tolerant
10. Imidazolinone tolerant	

Table 2 – Part 7. Non-limiting examples of traits/phenotypes with pest resistance properties

Legend

BR - Bacterial Resistant	NR - Nematode Resistant
FR - Fungal Resistant	VR - Viral Resistant
IR - Insect Resistant	

1. Agrobacterium resistant – BR	44. Ear mold resistant – FR
2. Alternaria resistant – FR	45. Erwinia carotovora resistant – BR
3. Alternaria daucii resistant – FR	46. European Corn Borer resistant – IR
4. Alternaria solani resistant – FR	47. Eyespot resistant – FR
5. AMV resistant – VR	48. Fall armyworm resistant – IR
6. Anthracnose resistant – FR	49. Fire blight resistant – BR
7. Aphid resistant – IR	50. Frogeye leaf spot resistant – FR
8. Apple scab resistant – FR	51. Fruit rot resistant – FR
9. Aspergillus resistant – FR	52. Fungal post-harvest resistant – FR
10. Bacterial leaf blight resistant – BR	53. Fungal resistant – FR
11. Bacterial resistant – BR	54. Fungal resistant general – FR
12. Bacterial soft rot resistant – BR	55. Fusarium deblae resistant – FR
13. Bacterial soft rot resistant – VR	56. Fusarium resistant – FR
14. Bacterial speck resistant – BR	57. Geminivirus resistant – VR
15. BCTV resistant – VR	58. Gray lead spot resistant – FR
16. Black shank resistant – FR	59. Helminthosporium resistant – FR
17. BLRV resistant – VR	60. Hordothionin produced – BR
18. BNYVV resistant – VR	61. Insect predator resistant – IR
19. Botrytis cinerea resistant – FR	62. Insect resistant general – IR
20. Botrytis resistant – FR	63. Late blight resistant – FR
21. BPMV resistant – VR	64. Leaf blight resistant – FR
22. Brown spot resistant – FR	65. Leaf spot resistant – FR
23. BYDV resistant – VR	66. Lepidopteran resistant – IR
24. BYMV resistant – VR	67. Lesser cornstalk borer resistant – IR
25. CaMV resistant – VR	68. LMV resistant – VR
26. Cercospora resistant – FR	69. Loss of systemic resistance – VR
27. Clavibacter resistant – BR	70. Marssonina resistant – FR
28. Closterovirus resistant – BR	71. MCDV resistant – VR
29. CLRV resistant – VR	72. MCMV resistant – VR
30. CMV resistant – FR	73. MDMV resistant – VR
31. Coleopteran resistant – IR	74. MDMV-B resistant – VR
32. Colletotrichum resistant – FR	75. Mealybug wilt virus resistant – VR
33. Colorado potato beetle resistant – IR	76. Melampsora resistant – FR
34. Corn earworm resistant – IR	77. Melodogyne resistant – NR
35. Corynebacterium sepedonicum resistant – BR	78. Meloidogyne resistant – NR
36. Cottonwood leaf beetle resistant – IR	79. Mexican Rice Borer resistant – IR
37. Cricconemella resistant – NR	80. Mycotoxin degradation – FR
38. Crown gall resistant – BR	81. Nepovirus resistant – VR
39. Cucumovirus resistant – VR	82. Northern corn leaf blight resistant – IR
40. Cylindrosporium resistant – FR	83. Nucleocapsid protein produced – VR
41. Disease resistant general – FR	84. Oblique banded leafhopper resistant – IR
42. Dollar spot resistant – FR	85. Oomycete resistant – FR
43. Downy mildew resistant – FR	86. Pathogenesis related proteins level increased – FR

Table 2 - Part 7. (continued) Non-limiting examples of traits/phenotypes with pest resistance properties

87. PEMV resistant - VR	116. SMV resistant - VR
88. PeSV Resistant - VR	117. Sod web worm resistant - IR
89. Phatara leaf beetle resistant - IR	118. Soft rot fungal resistant - FR
90. Phoma resistant - FR	119. Soft rot resistant - BR
91. Phytophthora resistant - FR	120. Southwestern corn borer resistant - IR
92. PLRV resistant - VR	121. SPFMV resistant - VR
93. Potyvirus resistant - VR	122. Sphaeropsis fruit rot resistant - FR
94. Powdery mildew resistant - FR	123. SqMV resistant - VR
95. PPV resistant - VR	124. SrMV resistant - VR
96. Pratylenchus vulnus resistant - NR	125. Streptomyces scabies resistant - BR
97. PRSV resistant - VR	126. Sugar cane borer resistant - IR
98. PRV resistant - VR	127. TEV resistant - VR
99. PSbMV resistant - VR	128. Thelaviopsis resistant - FR
100. Pseudomonas syringae resistant - BR	129. TMV resistant - FR
101. PSTV resistant - VR	130. Tobamovirus resistant - VR
102. PVX resistant - VR	131. ToMoV resistant - VR
103. PVY resistant - VR	132. ToMV resistant - VR
104. RBDV resistant - VR	133. TRV resistant - VR
105. Rhizoctonia resistant - FR	134. TSWV resistant - VR
106. Rhizoctonia solani resistant - FR	135. TVMV resistant - VR
107. Ring rot resistance - BR	136. TYLCV resistant - VR
108. Root-knot nematode resistant - NR	137. Venturia resistant - FR
109. Rust resistant - FR	138. Verticillium dahliae resistant - FR
110. SbMV resistant - VR	139. Verticillium resistant - FR
111. Sclerotinia resistant - FR	140. Western corn root worm resistant - IR
112. SCMV resistant - VR	141. WMV2 resistant - VR
113. SCYLV resistant - VR	142. WSMV resistant - VR
114. Septoria resistant - FR	143. ZYMV resistant - VR
115. Smut resistant - FR	

Table 2 - Part 8. Non-limiting examples of miscellaneous traits/phenotypes with properties

1. Antibiotic produced	31. Mycotoxin production inhibited
2. Antiprotease producing	32. Mycotoxin restored
3. Capable of growth on defined synthetic media	33. Non-lesion forming mutant
4. Carbohydrate metabolism altered	34. Novel protein produced
5. Cell wall altered	35. Oil quality altered
6. Cold tolerant	36. Peroxidase levels increased
7. Coleopteran resistant	37. Pharmaceutical proteins produced
8. Color altered	38. Phosphinothricin tolerant
9. Color sectors in seeds	39. Pigment metabolism altered
10. Colored sectors in leaves	40. Pollen visual marker
11. Constitutive expression of glutamine synthetase	41. Polyamine metabolism altered
12. Cre recombinase produced	42. Polymer produced
13. Dalapon tolerant	43. Recombinase produced
14. Development altered	44. Secondary metabolite increased
15. Disease resistant general	45. Seed color altered
16. Ethylene metabolism altered	46. Seed weight increased
17. Expression optimization	47. Selectable marker
18. Fenthion susceptible	48. Spectromycin resistant
19. Glucuronidase expressing	49. Sterile
20. Glyphosate tolerant	50. Sterols increased
21. Growth rate reduced	51. Sulfonylurea susceptible
22. Heavy metals sequestered	52. Syringomycin deficient
23. Hygromycin tolerant	53. Transposon activator
24. Inducible DNA modification	54. Transposon elements inserted
25. Industrial enzyme produced	55. Transposon inserted
26. Kanamycin resistant	56. Trifolotoxin producing
27. Lipase expressed in seeds	57. Trifolotoxin resistant

28. Methotrexate resistant	58. Virulence reduced
29. Modified growth characteristics	59. Visual marker
30. Mycotoxin deficient	60. Visual marker inactive

In a particular exemplification, "producing an organism having a desirable trait" includes an organism that is with respect to an organ or a part of an organ but not necessarily altered anywhere else.

By "trait" is meant any detectable parameter associated with an organism under a set of conditions. Examples of "detectable parameters" include the ability to produce a substance, the ability to not produce a substance, an altered pattern of (such as an increased or a decreased) ability to produce a substance, viability, non-viability, behaviour, growth rate, size, morphology or morphological characteristic,

In another embodiment, this invention is directed to a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining an initial population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one initial organism, c) selecting at least one mutagenized organism having a desirable trait or a desirable improvement in a trait, and d) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method. A mutagenized organism having a desirable trait or a desirable improvement in a trait can be referred to as an "up-mutant", and the associated mutation(s) contained in an up-mutant organism can be referred to as up-mutation(s).

In one embodiment, step c) is comprised of selecting at least two different mutagenized organisms, each having a different mutagenized genome, and the method of producing an organism having a desirable trait or a desirable improvement in a trait is comprised of a) obtaining a starting population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one starting organism, c) selecting at least two mutagenized organism having a desirable trait or a desirable improvement in a trait, d) creating combinations of the mutations of the two or more mutagenized organisms, e) selecting at least one mutagenized organism having a desirable trait or a desirable

improvement in a trait, and f) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method.

In one embodiment, the method is repeated. Thus, for example, an up-mutant organism can serve as a starting organism for the above method. Also, for example, an up mutant organism having a combination of two or more up-mutations in its genome can serve as a starting organism for the above method.

Thus, in one embodiment, this invention is directed to a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining a starting population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one starting organism, c) selecting at least one mutagenized organism having a desirable trait or a desirable improvement in a trait, and d) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method. A mutagenized organism having a desirable trait or a desirable improvement in a trait can be referred to as an "up-mutant", and the associated mutation(s) contained in an up-mutant organism can be referred to as up-mutation(s).

Mutagenizing a starting population such that mutations occur throughout a substantial part of the genome of at least one starting organism refers to mutagenizing at least approximately 1% of the genes of a genome, or at least approximately 10% of the genes of a genome, or at least approximately 20% of the genes of a genome, or at least approximately 30% of the genes of a genome, or at least approximately 40% of the genes of a genome, or at least approximately 50% of the genes of a genome, or at least approximately 60% of the genes of a genome, or at least approximately 70% of the genes of a genome, or at least approximately 80% of the genes of a genome, or at least approximately 90% of the genes of a genome, or at least approximately 95% of the genes of a genome, or at least approximately 98% of the genes of a genome.

In a particular embodiment, this invention provides a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining sequence information of a genome; b) annotating the genomic sequence obtained; c)

mutagenizing a substantial part of the genome the genome; d) selecting at least one mutagenized genome having a desirable trait or a desirable improvement in a trait; and e) optionally repeating the method by subjecting one or more mutagenized genomes to a repetition of the method.

Thus in one aspect, this invention provides a process comprised of:

- 1.) Subjecting a working cell or organism to holistic monitoring (which can include the detection and/or measurement of all detectable functions and physical parameters). Examples of such parameters include morphology, behavior, growth, responsiveness to stimuli (e.g., antibiotics, different environment, etc.). Additional examples include all measurable molecules, including molecules that are chemically at least in part a nucleic acids, proteins, carbohydrates, proteoglycans, glycoproteins, or lipids. In a particular aspect, performing holistic monitoring is comprised of using a microarray-based method. In another aspect, performing holistic monitoring is comprised of sequencing a substantial portion of the genome, i.e. for example at least approximately 10% of the genome, or for example at least approximately 20% of the genome, or for example at least approximately 30% of the genome, or for example at least approximately 40% of the genome, or for example at least approximately 50% of the genome, or for example at least approximately 60% of the genome, or for example at least approximately 70% of the genome, or for example at least approximately 80% of the genome, or for example at least approximately 90% of the genome, or for example at least approximately 95% of the genome, or for example at least approximately 98% of the genome.
- 2) Introducing into the working cell or organism a plurality of traits (stacked traits), including selectively and differentially activatable traits. Serviceable traits for this purpose include traits conferred by genes and traits conferred by gene pathways.
- 3) Subjecting the working cell or organism to holistic monitoring.
- 4) Compiling the information obtained from steps 1) and 3), and processing &/or analyzing it to better understand the changes introduced into the working cell or organisms. Such data processing includes identifying correlations between and/or among the measured parameters.

5) Repeating any number or all of steps 2), 3), and 4).

This invention provides that molecules serviceable for introducing transgenic traits into a plant include all known genes and nucleic acids. By way of non-limiting exemplification, this invention specifically names any number &/or combination of genes listed herein or listed in any reference incorporated herein by reference. Furthermore, by way of non-limiting exemplification, this invention specifically names any number &/or combination of genes & gene pathways listed herein as well as in any reference incorporated by reference herein. This invention provides that molecules serviceable as detectable parameters include molecule, any enzyme, substrate thereof, product thereof, and any gene or gene pathway listed herein including in any figure or table herein as well as in any reference incorporated by reference herein.

This invention also relates generally to the field of nucleic acid engineering and correspondingly encoded recombinant protein engineering. More particularly, the invention relates to the directed evolution of nucleic acids and screening of clones containing the evolved nucleic acids for resultant activity(ies) of interest, such nucleic acid activity(ies) &/or specified protein, particularly enzyme, activity(ies) of interest.

Mutagenized molecules provided by this invention may have chimeric molecules and molecules with point mutations, including biological molecules that contain a carbohydrate, a lipid, a nucleic acid, &/or a protein component, and specific but non-limiting examples of these include antibiotics, antibodies, enzymes, and steroidal and non-steroidal hormones.

This invention relates generally to a method of: 1) preparing a progeny generation of molecule(s) (including a molecule that is comprised of a polynucleotide sequence, a molecule that is comprised of a polypeptide sequence, and a molecules that is comprised in part of a polynucleotide sequence and in part of a polypeptide sequence), that is mutagenized to achieve at least one point mutation, addition, deletion, &/or chimerization, from one or more ancestral or parental generation template(s); 2) screening the progeny generation molecule(s) - preferably using a high throughput method - for at least one property of interest (such as an improvement in an enzyme activity or an increase in stability or a novel chemotherapeutic

effect); 3) optionally obtaining &/or cataloguing structural &/or and functional information regarding the parental &/or progeny generation molecules; and 4) optionally repeating any of steps 1) to 3).

In a preferred embodiment, there is generated (e.g. from a parent polynucleotide template) - in what is termed "codon site-saturation mutagenesis" - a progeny generation of polynucleotides, each having at least one set of up to three contiguous point mutations (i.e. different bases comprising a new codon), such that every codon (or every family of degenerate codons encoding the same amino acid) is represented at each codon position. Corresponding to - and encoded by - this progeny generation of polynucleotides, there is also generated a set of progeny polypeptides, each having at least one single amino acid point mutation. In a preferred aspect, there is generated - in what is termed "amino acid site-saturation mutagenesis" - one such mutant polypeptide for each of the 19 naturally encoded polypeptide-forming alpha-amino acid substitutions at each and every amino acid position along the polypeptide. This yields - for each and every amino acid position along the parental polypeptide - a total of 20 distinct progeny polypeptides including the original amino acid, or potentially more than 21 distinct progeny polypeptides if additional amino acids are used either instead of or in addition to the 20 naturally encoded amino acids

Thus, in another aspect, this approach is also serviceable for generating mutants containing - in addition to &/or in combination with the 20 naturally encoded polypeptide-forming alpha-amino acids - other rare &/or not naturally-encoded amino acids and amino acid derivatives. In yet another aspect, this approach is also serviceable for generating mutants by the use of - in addition to &/or in combination with natural or unaltered codon recognition systems of suitable hosts - altered, mutagenized, &/or designer codon recognition systems (such as in a host cell with one or more altered tRNA molecules).

In yet another aspect, this invention relates to recombination and more specifically to a method for preparing polynucleotides encoding a polypeptide by a method of *in vivo* re-assortment of polynucleotide sequences containing regions of partial homology, assembling the polynucleotides to form at least one polynucleotide and screening the polynucleotides for the production of polypeptide(s) having a useful property.

In yet another preferred embodiment, this invention is serviceable for analyzing and cataloguing - with respect to any molecular property (e.g. an enzymatic activity) or combination of properties allowed by current technology - the effects of any mutational change achieved (including particularly saturation mutagenesis). Thus, a comprehensive method is provided for determining the effect of changing each amino acid in a parental polypeptide into each of at least 19 possible substitutions. This allows each amino acid in a parental polypeptide to be characterized and catalogued according to its spectrum of potential effects on a measurable property of the polypeptide.

In another aspect, the method of the present invention utilizes the natural property of cells to recombine molecules and/or to mediate reductive processes that reduce the complexity of sequences and extent of repeated or consecutive sequences possessing regions of homology.

It is an object of the present invention to provide a method for generating hybrid polynucleotides encoding biologically active hybrid polypeptides with enhanced activities. In accomplishing these and other objects, there has been provided, in accordance with one aspect of the invention, a method for introducing polynucleotides into a suitable host cell and growing the host cell under conditions that produce a hybrid polynucleotide.

In another aspect of the invention, the invention provides a method for screening for biologically active hybrid polypeptides encoded by hybrid polynucleotides. The present method allows for the identification of biologically active hybrid polypeptides with enhanced biological activities.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

In yet another aspect, this invention relates to a method of discovering which phenotype corresponds to a gene by disrupting every gene in the organism.

Accordingly, this invention provides a method for determining a gene that alters a characteristic of an organism, comprising: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence an organism having an altered trait, and d) determining the nucleotide sequence of a gene that has been mutagenized in the organism having the altered trait.

In yet another aspect, this invention relates to a method of improving a trait in an organism by functionally knocking out a particular gene in the organism, and then transferring a library of genes, which only vary from the wild-type at one codon position, into the organism.

Accordingly, this invention provides a method method for producing an organism with an improved trait, comprising:

- a) functionally knocking out an enogenous gene in a substantially clonal population of organisms;
- b) transferring the set of altered genes into the clonal population of organisms, wherein each altered gene differs from the endogenous gene at only one codon; and
- c) detecting a mutagenized organism having an improved trait; and
- d) determining the nucleotide sequence of a gene that has been transferred into the detected organism.

D. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Exonuclease Activity. Figure 1 shows the activity of the enzyme exonuclease III. This is an exemplary enzyme that can be used to shuffle, assemble, reassemble, recombine, and/or concatenate polynucleotide building blocks. The asterisk indicates that the enzyme acts from the 3' direction towards the 5' direction of the polynucleotide substrate.

Figure 2. Generation of A Nucleic Acid Building Block by Polymerase-Based Amplification. Figure 2 illustrates a method of generating a double-stranded nucleic acid building block with two overhangs using a polymerase-based amplification reaction (e.g., PCR). As illustrated, a first polymerase-based amplification reaction using a first set of primers, F_2 and R_1 , is used to generate a blunt-ended product (labeled Reaction 1, Product 1), which is essentially identical to Product A. A second polymerase-based amplification reaction using a second set of primers, F_1 and R_2 , is used to generate a blunt-ended product (labeled Reaction 2, Product 2), which is essentially identical to Product B. These two products are then mixed and allowed to melt and anneal, generating a potentially useful double-stranded nucleic acid building block with two overhangs. In the example of Fig. 1, the product with the 3' overhangs (Product C) is selected for by nuclease-based degradation of the other 3 products using a 3' acting exonuclease, such as exonuclease III. Alternate primers are shown in parenthesis to illustrate serviceable primers may overlap, and additionally that serviceable primers may be of different lengths, as shown.

FIGURE 3. Unique Overhangs And Unique Couplings. Figure 3 illustrates the point that the number of unique overhangs of each size (e.g. the total number of unique overhangs composed of 1 or 2 or 3, etc. nucleotides) exceeds the number of unique couplings that can result from the use of all the unique overhangs of that size. For example, there are 4 unique 3' overhangs composed of a single nucleotide, and 4 unique 5' overhangs composed of a single nucleotide. Yet the total number of unique couplings that can be made using all the 8 unique single-nucleotide 3' overhangs and single-nucleotide 5' overhangs is 4.

FIGURE 4. Unique Overall Assembly Order Achieved by Sequentially Coupling the Building Blocks

Figure 4 illustrates the fact that in order to assemble a total of "n" nucleic acid building blocks, "n-1" couplings are needed. Yet it is sometimes the case that the number of unique couplings available for use is fewer than the "n-1" value. Under these, and other, circumstances a stringent non-stochastic overall assembly order can still be achieved by performing the assembly process in sequential steps. In this example, 2 sequential steps are used to achieve a designed overall assembly order for five nucleic acid building

blocks. In this illustration the designed overall assembly order for the five nucleic acid building blocks is: 5'-(#1-#2-#3-#4-#5)-3', where #1 represents building block number 1, etc.

FIGURE 5. Unique Couplings Available Using a Two-Nucleotide 3' Overhang.

Figure 5 further illustrates the point that the number of unique overhangs of each size (here, e.g. the total number of unique overhangs composed of 2 nucleotides) exceeds the number of unique couplings that can result from the use of all the unique overhangs of that size. For example, there are 16 unique 3' overhangs composed of two nucleotides, and another 16 unique 5' overhangs composed of two nucleotides, for a total of 32 as shown. Yet the total number of couplings that are unique and not self-binding that can be made using all the 32 unique double-nucleotide 3' overhangs and double-nucleotide 5' overhangs is 12. Some apparently unique couplings have "identical twins" (marked in the same shading), which are visually obvious in this illustration. Still other overhangs contain nucleotide sequences that can self-bind in a palindromic fashion, as shown and labeled in this figure; thus they not contribute the high stringency to the overall assembly order.

Figure 6. Generation of an Exhaustive Set of Chimeric Combinations by Synthetic Ligation Reassembly. Figure 6 showcases the power of this invention in its ability to generate exhaustively and systematically all possible combinations of the nucleic acid building blocks designed in this example. Particularly large sets (or libraries) of progeny chimeric molecules can be generated. Because this method can be performed exhaustively and systematically, the method application can be repeated by choosing new demarcation points and with correspondingly newly designed nucleic acid building blocks, bypassing the burden of re-generating and re-screening previously examined and rejected molecular species. It is appreciated that, codon wobble can be used to advantage to increase the frequency of a demarcation point. In other words, a particular base can often be substituted into a nucleic acid building block without altering the amino acid encoded by progenitor codon (that is now altered codon) because of codon degeneracy. As illustrated, demarcation points are chosen upon alignment of 8 progenitor templates. Nucleic acid building blocks including their overhangs (which are serviceable for the formation of ordered couplings) are then designed and synthesized. In this instance, 18 nucleic acid

building blocks are generated based on the sequence of each of the 8 progenitor templates, for a total of 144 nucleic acid building blocks (or double-stranded oligos). Performing the ligation synthesis procedure will then produce a library of progeny molecules comprised of yield of 8^{18} (or over 1.8×10^{16}) chimeras.

Figure 7. Synthetic genes from oligos:. According to one embodiment of this invention, double-stranded nucleic acid building blocks are designed by aligning a plurality of progenitor nucleic acid templates. Preferably these templates contain some homology and some heterology. The nucleic acids may encode related proteins, such as related enzymes, which relationship may be based on function or structure or both. Figure 7 shows the alignment of three polynucleotide progenitor templates and the selection of demarcation points (boxed) shared by all the progenitor molecules. In this particular example, the nucleic acid building blocks derived from each of the progenitor templates were chosen to be approximately 30 to 50 nucleotides in length.

Figure 8. Nucleic acid building blocks for synthetic ligation gene reassembly.

Figure 8 shows the nucleic acid building blocks from the example in Figure 7. The nucleic acid building blocks are shown here in generic cartoon form, with their compatible overhangs, including both 5' and 3' overhangs. There are 22 total nucleic acid building blocks derived from each of the 3 progenitor templates. Thus, the ligation synthesis procedure can produce a library of progeny molecules comprised of yield of 3^{22} (or over 3.1×10^{10}) chimeras.

Figure 9. Addition of Introns by Synthetic Ligation Reassembly. Figure 9 shows in generic cartoon form that an intron may be introduced into a chimeric progeny molecule by way of a nucleic acid building block. It is appreciated that introns often have consensus sequences at both termini in order to render them operational. It is also appreciated that, in addition to enabling gene splicing, introns may serve an additional purpose by providing sites of homology to other nucleic acids to enable homologous recombination. For this purpose, and potentially others, it may be sometimes desirable to generate a large nucleic acid building block for introducing an intron. If the size is overly large easily generating by direct chemical synthesis of two single stranded oligos, such a specialized nucleic acid building block may also be generated by direct chemical synthesis

of more than two single stranded oligos or by using a polymerase-based amplification reaction as shown in Figure 2.

Figure 10. Ligation Reassembly Using Fewer Than All The Nucleotides Of An Overhang. Figure 10 shows that coupling can occur in a manner that does not make use of every nucleotide in a participating overhang. The coupling is particularly lively to survive (e.g. in a transformed host) if the coupling reinforced by treatment with a ligase enzyme to form what may be referred to as a "gap ligation" or a "gapped ligation". It is appreciated that, as shown, this type of coupling can contribute to generation of unwanted background product(s), but it can also be used advantageously increase the diversity of the progeny library generated by the designed ligation reassembly.

Figure 11. Avoidance of unwanted self-ligation in palindromic couplings. As mentioned before and shown in Figure 5, certain overhangs are able to undergo self-coupling to form a palindromic coupling. A coupling is strengthened substantially if it is reinforced by treatment with a ligase enzyme. Accordingly, it is appreciated that the lack of 5' phosphates on these overhangs, as shown, can be used advantageously to prevent this type of palindromic self-ligation. Accordingly, this invention provides that nucleic acid building blocks can be chemically made (or ordered) that lack a 5' phosphate group (or alternatively they can be remove – e.g. by treatment with a phosphatase enzyme such as a calf intestinal alkaline phosphatase (CIAP) – in order to prevent palindromic self-ligations in ligation reassembly processes.

Figure 12. Pathway Engineering. It is a goal of this invention to provide ways of making new gene pathways using ligation reassembly, optionally with other directed evolution methods such as saturation mutagenesis. Figure 12 illustrates a preferred approach that may be taken to achieve this goal. It is appreciated that naturally-occurring microbial gene pathways are linked more often than naturally-occurring eukaryotic (e.g. plant) gene pathways, which are sometime only partially linked. In a particular embodiment, this invention provides that regulatory gene sequences (including promoters) can be introduced in the form of nucleic acid building blocks into progeny gene pathways generated by ligation reassembly processes. Thus, originally linked microbial gene

pathways, as well as originally unlinked genes and gene pathways, can be thus converted to acquire operability in plants and other eukaryotes.

Figure 13. Avoidance of unwanted self-ligation in palindromic couplings. Figure 13 illustrates that another goal of this invention, in addition to the generation of novel gene pathways, is the subjection of gene pathways – both naturally occurring and man-made – to mutagenesis and selection in order to achieve improved progeny molecules using the instantly disclosed methods of directed evolution (including saturation mutagenesis and synthetic ligation reassembly). In a particular embodiment, as provided by the instant invention, both microbial and plant pathways can be improved by directed evolution, and as shown, the directed evolution process can be performed both on genes prior to linking them into pathways, and on gene pathways themselves.

Figure 14. Conversion of Microbial Pathways to Eukaryotic Pathways. In a particular embodiment, this invention provides that microbial pathways can be converted to pathways operable in plants and other eukaryotic species by the introduction of regulatory sequences that function in those species. Preferred regulatory sequences include promoters, operators, and activator binding sites. As shown, a preferred method of achieving the introduction of such serviceable regulatory sequences is in the form of nucleic acid building blocks, particularly through the use of couplings in ligation reassembly processes. These couplings in Fig. 14 are marked with the letters A, B, C, D and F.

Fig. 15. Holistic engineering of differentially activatable stacked traits in noveltransgenic plants using directed evolution and whole cell monitoring.

Fig. 16. Differential Activation of Selected Traits Can Be Achieved by Adjusting and Controlling the Environment of the Traits.

Fig. 17. Harvesting, Processing, Storage.

Fig. 18. Processing.

Fig. 19. Cellular Mutagenesis.

Figure 20. Differential Activation of Selected Precursor (Inactive) Gene Products.

Figure 21. Starting population comprised of an organism strain to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved organism strain that has a desired trait.

Figure 22. Starting population comprised of a genomic sequence to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved genomic sequence that has a desired trait.

Fig. 23. Strain Improvement.

Fig. 24. Iterative Strain Improvement.

E. DEFINITIONS OF TERMS

In order to facilitate understanding of the examples provided herein, certain frequently occurring methods and/or terms will be described.

The term "agent" is used herein to denote a chemical compound, a mixture of chemical compounds, an array of spatially localized compounds (e.g., a VLSIPS peptide array, polynucleotide array, and/or combinatorial small molecule array), biological macromolecule, a bacteriophage peptide display library, a bacteriophage antibody (e.g., scFv) display library, a polysome peptide display library, or an extract made from biological materials such as bacteria, plants, fungi, or animal (particular mammalian) cells or tissues. Agents are evaluated for potential activity as anti-neoplastics, anti-inflammatories or apoptosis modulators by inclusion in screening assays described hereinbelow. Agents are evaluated for potential activity as specific protein interaction inhibitors (i.e., an agent which selectively inhibits a binding interaction between two predetermined polypeptides but which does not substantially interfere with cell viability) by inclusion in screening assays described hereinbelow.

An "ambiguous base requirement" in a restriction site refers to a nucleotide base requirement that is not specified to the fullest extent, i.e. that is not a specific base (such as, in a non-limiting exemplification, a specific base selected from A, C, G, and T), but rather may be any one of at least two or more bases. Commonly accepted abbreviations that are used in the art as well as herein to represent ambiguity in bases include the following: R = G or A; Y = C or T; M = A or C; K = G or T; S = G or C; W = A or T; H = A or C or T; B = G or T or C; V = G or C or A; D = G or A or T; N = A or C or G or T.

The term "amino acid" as used herein refers to any organic compound that contains an amino group (-NH₂) and a carboxyl group (-COOH); preferably either as free groups or alternatively after condensation as part of peptide bonds. The "twenty naturally encoded polypeptide-forming alpha-amino acids" are understood in the art and refer to: alanine (ala or A), arginine (arg or R), asparagine (asn or N), aspartic acid (asp or D), cysteine (cys or C), glutamic acid (glu or E), glutamine (gln or Q), glycine (gly or G), histidine (his or H), isoleucine (ile or I), leucine (leu or L), lysine (lys or K), methionine (met or M), phenylalanine (phe or F), proline (pro or P), serine (ser or S), threonine (thr or T), tryptophan (trp or W), tyrosine (tyr or Y), and valine (val or V).

The term "amplification" means that the number of copies of a polynucleotide is increased.

The term "antibody", as used herein, refers to intact immunoglobulin molecules, as well as fragments of immunoglobulin molecules, such as Fab, Fab', (Fab')₂, Fv, and SCA fragments, that are capable of binding to an epitope of an antigen. These antibody fragments, which retain some ability to selectively bind to an antigen (e.g., a polypeptide antigen) of the antibody from which they are derived, can be made using well known methods in the art (see, e.g., Harlow and Lane, *supra*), and are described further, as follows.

- (1) An Fab fragment consists of a monovalent antigen-binding fragment of an antibody molecule, and can be produced by digestion of a whole antibody

molecule with the enzyme papain, to yield a fragment consisting of an intact light chain and a portion of a heavy chain.

- (2) An Fab' fragment of an antibody molecule can be obtained by treating a whole antibody molecule with pepsin, followed by reduction, to yield a molecule consisting of an intact light chain and a portion of a heavy chain. Two Fab' fragments are obtained per antibody molecule treated in this manner.
- (3) An (Fab')₂ fragment of an antibody can be obtained by treating a whole antibody molecule with the enzyme pepsin, without subsequent reduction. A (Fab')₂ fragment is a dimer of two Fab' fragments, held together by two disulfide bonds.
- (4) An Fv fragment is defined as a genetically engineered fragment containing the variable region of a light chain and the variable region of a heavy chain expressed as two chains.
- (5) An single chain antibody ("SCA") is a genetically engineered single chain molecule containing the variable region of a light chain and the variable region of a heavy chain, linked by a suitable, flexible polypeptide linker.

The term "Applied Molecular Evolution" ("AME") means the application of an evolutionary design algorithm to a specific, useful goal. While many different library formats for AME have been reported for polynucleotides, peptides and proteins (phage, lacI and polysomes), none of these formats have provided for recombination by random cross-overs to deliberately create a combinatorial library.

A molecule that has a "chimeric property" is a molecule that is: 1) in part homologous and in part heterologous to a first reference molecule; while 2) at the same time being in part homologous and in part heterologous to a second reference molecule; without 3) precluding the possibility of being at the same time in part homologous and in part heterologous to still one or more additional reference molecules. In a non-limiting embodiment, a chimeric molecule may be prepared by assembling a reassortment of

partial molecular sequences. In a non-limiting aspect, a chimeric polynucleotide molecule may be prepared by synthesizing the chimeric polynucleotide using plurality of molecular templates, such that the resultant chimeric polynucleotide has properties of a plurality of templates.

The term "cognate" as used herein refers to a gene sequence that is evolutionarily and functionally related between species. For example, but not limitation, in the human genome the human CD4 gene is the cognate gene to the mouse 3d4 gene, since the sequences and structures of these two genes indicate that they are highly homologous and both genes encode a protein which functions in signaling T cell activation through MHC class II-restricted antigen recognition.

A "comparison window," as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith (Smith and Waterman, *Adv Appl Math*, 1981; Smith and Waterman, *J Teor Biol*, 1981; Smith and Waterman, *J Mol Biol*, 1981; Smith et al, *J Mol Evol*, 1981), by the homology alignment algorithm of Needleman (Needleman and Wuncsch, 1970), by the search of similarity method of Pearson (Pearson and Lipman, 1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection, and the best alignment (i.e., resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected.

As used herein, the term "complementarity-determining region" and "CDR" refer to the art-recognized term as exemplified by the Kabat and Chothia CDR definitions also generally known as supervariable regions or hypervariable loops (Chothia and Lesk, 1987; Chothia et al, 1989; Kabat et al, 1987; and Tramontano et al, 1990). Variable

region domains typically comprise the amino-terminal approximately 105-115 amino acids of a naturally-occurring immunoglobulin chain (e.g., amino acids 1-110), although variable domains somewhat shorter or longer are also suitable for forming single-chain antibodies.

"Conservative amino acid substitutions" refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are : valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

The term **"corresponds to"** is used herein to mean that a polynucleotide sequence is homologous (i.e., is identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term **"complementary to"** is used herein to mean that the complementary sequence is homologous to all or a portion of a reference polynucleotide sequence. For illustration, the nucleotide sequence **"TATAC"** corresponds to a reference **"TATAC"** and is complementary to a reference sequence **"GTATA."**

The term **"degrading effective"** amount refers to the amount of enzyme which is required to process at least 50% of the substrate, as compared to substrate not contacted with the enzyme. Preferably, at least 80% of the substrate is degraded.

As used herein, the term **"defined sequence framework"** refers to a set of defined sequences that are selected on a non-random basis, generally on the basis of experimental data or structural data; for example, a defined sequence framework may comprise a set of

amino acid sequences that are predicted to form a β -sheet structure or may comprise a leucine zipper heptad repeat motif, a zinc-finger domain, among other variations. A "defined sequence kernel" is a set of sequences which encompass a limited scope of variability. Whereas (1) a completely random 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences, and (2) a pseudorandom 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences but will exhibit a bias for certain residues at certain positions and/or overall, (3) a defined sequence kernel is a subset of sequences if each residue position was allowed to be any of the allowable 20 conventional amino acids (and/or allowable unconventional amino/imino acids). A defined sequence kernel generally comprises variant and invariant residue positions and/or comprises variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), and the like, either segmentally or over the entire length of the individual selected library member sequence. Defined sequence kernels can refer to either amino acid sequences or polynucleotide sequences. Of illustration and not limitation, the sequences $(\text{NNK})_{10}$ and $(\text{NNM})_{10}$, wherein N represents A, T, G, or C; K represents G or T; and M represents A or C, are defined sequence kernels.

"Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 μg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 μl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the reaction is electrophoresed directly on a gel to isolate the desired fragment.

"Directional ligation" refers to a ligation in which a 5' end and a 3' end of a polynucleotide are different enough to specify a preferred ligation orientation. For example, an otherwise untreated and undigested PCR product that has two blunt ends will typically not have a preferred ligation orientation when ligated into a cloning vector

digested to produce blunt ends in its multiple cloning site; thus, directional ligation will typically not be displayed under these circumstances. In contrast, directional ligation will typically displayed when a digested PCR product having a 5' *EcoR* I-treated end and a 3' *BamH* I-is ligated into a cloning vector that has a multiple cloning site digested with *EcoR* I and *BamH* I.

The term "DNA shuffling" is used herein to indicate recombination between substantially homologous but non-identical sequences, in some embodiments DNA shuffling may involve crossover via non-homologous recombination, such as via *cer/lox* and/or *flp/fit* systems and the like.

As used in this invention, the term "epitope" refers to an antigenic determinant on an antigen, such as a phytase polypeptide, to which the paratope of an antibody, such as an phytase-specific antibody, binds. Antigenic determinants usually consist of chemically active surface groupings of molecules, such as amino acids or sugar side chains, and can have specific three-dimensional structural characteristics, as well as specific charge characteristics. As used herein "epitope" refers to that portion of an antigen or other macromolecule capable of forming a binding interaction that interacts with the variable region binding body of an antibody. Typically, such binding interaction is manifested as an intermolecular contact with one or more amino acid residues of a CDR.

The terms "fragment", "derivative" and "analog" when referring to a reference polypeptide comprise a polypeptide which retains at least one biological function or activity that is at least essentially same as that of the reference polypeptide. Furthermore, the terms "fragment", "derivative" or "analog" are exemplified by a "pro-form" molecule, such as a low activity proprotein that can be modified by cleavage to produce a mature enzyme with significantly higher activity.

A method is provided herein for producing from a template polypeptide a set of progeny polypeptides in which a "full range of single amino acid substitutions" is represented at each amino acid position. As used herein, "full range of single amino acid substitutions" is in reference to the naturally encoded 20 naturally encoded polypeptide-forming alpha-amino acids, as described herein.

The term **"gene"** means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

"Genetic instability", as used herein, refers to the natural tendency of highly repetitive sequences to be lost through a process of reductive events generally involving sequence simplification through the loss of repeated sequences. Deletions tend to involve the loss of one copy of a repeat and everything between the repeats.

The term **"heterologous"** means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that areas of polynucleotides or polynucleotides have areas or regions within their sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are for example areas of mutations.

The term **"homologous"** or **"homeologous"** means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentrations as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp.

An immunoglobulin light or heavy chain variable region consists of a **"framework"** region interrupted by three hypervariable regions, also called CDR's. The extent of the framework region and CDR's have been precisely defined; see *"Sequences of Proteins of Immunological Interest"* (Kabat et al, 1987). The sequences of the framework regions of different light or heavy chains are relatively conserved within a specie. As used herein, a **"human framework region"** is a framework region that is substantially identical (about 85 or more, usually 90-95 or more) to the framework region of a naturally occurring human immunoglobulin. the framework region of an antibody, that is the combined framework regions of the constituent light and heavy chains, serves to position and align the CDR's. The CDR's are primarily responsible for binding to an epitope of an antigen.

The benefits of this invention extend to "commercial applications" (or commercial processes), which term is used to include applications in commercial industry proper (or simply industry) as well as non-commercial commercial applications (e.g. biomedical research at a non-profit institution). Relevant applications include those in areas of diagnosis, medicine, agriculture, manufacturing, and academia.

The term "identical" or "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a polynucleotide or the overall polynucleotide are identical or complementary to areas of another polynucleotide or the polynucleotide.

The term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or enzyme present in a living animal is not isolated, but the same polynucleotide or enzyme, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or enzymes could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

By "isolated nucleic acid" is meant a nucleic acid, e.g., a DNA or RNA molecule, that is not immediately contiguous with the 5' and 3' flanking sequences with which it normally is immediately contiguous when present in the naturally occurring genome of the organism from which it is derived. The term thus describes, for example, a nucleic acid that is incorporated into a vector, such as a plasmid or viral vector; a nucleic acid that is incorporated into the genome of a heterologous cell (or the genome of a homologous cell, but at a site different from that at which it naturally occurs); and a nucleic acid that exists as a separate molecule, e.g., a DNA fragment produced by PCR amplification or restriction enzyme digestion, or an RNA molecule produced by *in vitro* transcription. The term also describes a recombinant nucleic acid that forms part of a hybrid gene encoding additional polypeptide sequences that can be used, for example, in the production of a fusion protein.

As used herein "ligand" refers to a molecule, such as a random peptide or variable segment sequence, that is recognized by a particular receptor. As one of skill in the art will recognize, a molecule (or macromolecular complex) can be both a receptor and a ligand. In general, the binding partner having a smaller molecular weight is referred to as the ligand and the binding partner having a greater molecular weight is referred to as a receptor.

"Ligation" refers to the process of forming phosphodiester bonds between two double stranded nucleic acid fragments (Sambrook et al, 1982, p. 146; Sambrook, 1989). Unless otherwise provided, ligation may be accomplished using known buffers and conditions with 10 units of T4 DNA ligase ("ligase") per 0.5 µg of approximately equimolar amounts of the DNA fragments to be ligated.

As used herein, "linker" or "spacer" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide, and serves to place the two molecules in a preferred configuration, e.g., so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

As used herein, a "molecular property to be evolved" includes reference to molecules comprised of a polynucleotide sequence, molecules comprised of a polypeptide sequence, and molecules comprised in part of a polynucleotide sequence and in part of a polypeptide sequence. Particularly relevant - but by no means limiting - examples of molecular properties to be evolved include enzymatic activities at specified conditions, such as related to temperature; salinity; pressure; pH; and concentration of glycerol, DMSO, detergent, &/or any other molecular species with which contact is made in a reaction environment. Additional particularly relevant - but by no means limiting - examples of molecular properties to be evolved include stabilities - e.g. the amount of a residual molecular property that is present after a specified exposure time to a specified environment, such as may be encountered during storage.

The term "mutations" includes changes in the sequence of a wild-type or parental nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be

point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications. A mutation can also be a "**chimerization**", which is exemplified in a progeny molecule that is generated to contain part or all of a sequence of one parental molecule as well as part or all of a sequence of at least one other parental molecule. This invention provides for both chimeric polynucleotides and chimeric polypeptides.

As used herein, the degenerate "**N,N,G/T**" nucleotide sequence represents 32 possible triplets, where "**N**" can be A, C, G or T.

The term "**naturally-occurring**" as used herein as applied to the object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally occurring. Generally, the term naturally occurring refers to an object as present in a non-pathological (un-diseased) individual, such as would be typical for the species.

As used herein, a "**nucleic acid molecule**" is comprised of at least one base or one base pair, depending on whether it is single-stranded or double-stranded, respectively. Furthermore, a nucleic acid molecule may belong exclusively or chimerically to any group of nucleotide-containing molecules, as exemplified by, but not limited to, the following groups of nucleic acid molecules: RNA, DNA, genomic nucleic acids, non-genomic nucleic acids, naturally occurring and not naturally occurring nucleic acids, and synthetic nucleic acids. This includes, by way of non-limiting example, nucleic acids associated with any organelle, such as the mitochondria, ribosomal RNA, and nucleic acid molecules comprised chimerically of one or more components that are not naturally occurring along with naturally occurring components.

Additionally, a "**nucleic acid molecule**" may contain in part one or more non-nucleotide-based components as exemplified by, but not limited to, amino acids and sugars. Thus, by way of example, but not limitation, a ribozyme that is in part nucleotide-based and in part protein-based is considered a "**nucleic acid molecule**".

In addition, by way of example, but not limitation, a nucleic acid molecule that is labeled with a detectable moiety, such as a radioactive or alternatively a non-radioactive label, is likewise considered a **"nucleic acid molecule"**.

The terms **"nucleic acid sequence coding for"** or a **"DNA coding sequence of"** or a **"nucleotide sequence encoding"** a particular enzyme – as well as other synonymous terms – refer to a DNA sequence which is transcribed and translated into an enzyme when placed under the control of appropriate regulatory sequences. A **"promotor sequence"** is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. The promoter is part of the DNA sequence. This sequence region has a start codon at its 3' terminus. The promoter sequence does include the minimum number of bases where elements necessary to initiate transcription at levels detectable above background. However, after the RNA polymerase binds the sequence and transcription is initiated at the start codon (3' terminus with a promoter), transcription proceeds downstream in the 3' direction. Within the promoter sequence will be found a transcription initiation site (conveniently defined by mapping with nuclease S1) as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

The terms **"nucleic acid encoding an enzyme (protein)"** or **"DNA encoding an enzyme (protein)"** or **"polynucleotide encoding an enzyme (protein)"** and other synonymous terms encompasses a polynucleotide which includes only coding sequence for the enzyme as well as a polynucleotide which includes additional coding and/or non-coding sequence.

In one preferred embodiment, a **"specific nucleic acid molecule species"** is defined by its chemical structure, as exemplified by, but not limited to, its primary sequence. In another preferred embodiment, a specific **"nucleic acid molecule species"** is defined by a function of the nucleic acid species or by a function of a product derived from the nucleic acid species. Thus, by way of non-limiting example, a **"specific nucleic acid molecule species"** may be defined by one or more activities or properties attributable to it, including activities or properties attributable its expressed product.

The instant definition of “assembling a working nucleic acid sample into a nucleic acid library” includes the process of incorporating a nucleic acid sample into a vector-based collection, such as by ligation into a vector and transformation of a host. A description of relevant vectors, hosts, and other reagents as well as specific non-limiting examples thereof are provided hereinafter. The instant definition of “assembling a working nucleic acid sample into a nucleic acid library” also includes the process of incorporating a nucleic acid sample into a non-vector-based collection, such as by ligation to adaptors. Preferably the adaptors can anneal to PCR primers to facilitate amplification by PCR.

Accordingly, in a non-limiting embodiment, a “nucleic acid library” is comprised of a vector-based collection of one or more nucleic acid molecules. In another preferred embodiment a “nucleic acid library” is comprised of a non-vector-based collection of nucleic acid molecules. In yet another preferred embodiment a “nucleic acid library” is comprised of a combined collection of nucleic acid molecules that is in part vector-based and in part non-vector-based. Preferably, the collection of molecules comprising a library is searchable and separable according to individual nucleic acid molecule species.

The present invention provides a “nucleic acid construct” or alternatively a “nucleotide construct” or alternatively a “DNA construct”. The term “construct” is used herein to describe a molecule, such as a polynucleotide (*e.g.*, a phytase polynucleotide) may optionally be chemically bonded to one or more additional molecular moieties, such as a vector, or parts of a vector. In a specific - but by no means limiting - aspect, a nucleotide construct is exemplified by a DNA expression DNA expression constructs suitable for the transformation of a host cell.

An “oligonucleotide” (or synonymously an “oligo”) refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides may or may not have a 5' phosphate. Those that do not will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated. To achieve polymerase-based

amplification (such as with PCR), a **"32-fold degenerate oligonucleotide that is comprised of, in series, at least a first homologous sequence, a degenerate N,N,G/T sequence, and a second homologous sequence"** is mentioned. As used in this context, "homologous" is in reference to homology between the oligo and the parental polynucleotide that is subjected to the polymerase-based amplification.

As used herein, the term **"operably linked"** refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

A coding sequence is **"operably linked to"** another coding sequence when RNA polymerase will transcribe the two coding sequences into a single mRNA, which is then translated into a single polypeptide having amino acids derived from both coding sequences. The coding sequences need not be contiguous to one another so long as the expressed sequences are ultimately processed to produce the desired protein.

As used herein the term **"parental polynucleotide set"** is a set comprised of one or more distinct polynucleotide species. Usually this term is used in reference to a progeny polynucleotide set which is preferably obtained by mutagenization of the parental set, in which case the terms **"parental"**, **"starting"** and **"template"** are used interchangeably.

As used herein the term **"physiological conditions"** refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45 C and 0.001-10 mM divalent cation (e.g., Mg^{++} , Ca^{++}); preferably

about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v).

Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or non-ionic detergents and/or membrane fractions and/or anti-foam agents and/or scintillants.

Standard convention (5' to 3') is used herein to describe the sequence of double stranded polynucleotides.

The term "**population**" as used herein means a collection of components such as polynucleotides, portions or polynucleotides or proteins. A "**mixed population**:" means a collection of components which belong to the same family of nucleic acids or proteins (i.e., are related) but which differ in their sequence (i.e., are not identical) and hence in their biological activity.

A molecule having a "**pro-form**" refers to a molecule that undergoes any combination of one or more covalent and noncovalent chemical modifications (e.g. glycosylation, proteolytic cleavage, dimerization or oligomerization, temperature-induced or pH-induced conformational change, association with a co-factor, etc.) en route to attain a more mature molecular form having a property difference (e.g. an increase in activity) in comparison with the reference pro-form molecule. When two or more chemical modification (e.g. two proteolytic cleavages, or a proteolytic cleavage and a deglycosylation) can be distinguished en route to the production of a mature molecule, the reference precursor molecule may be termed a "**pre-pro-form**" molecule.

As used herein, the term "**pseudorandom**" refers to a set of sequences that have limited variability, such that, for example, the degree of residue variability at another position, but any pseudorandom position is allowed some degree of residue variation, however circumscribed.

"Quasi-repeated units", as used herein, refers to the repeats to be re-assorted and are by definition not identical. Indeed the method is proposed not only for practically identical encoding units produced by mutagenesis of the identical starting sequence, but also the reassortment of similar or related sequences which may diverge significantly in some regions. Nevertheless, if the sequences contain sufficient homologies to be reassorted by this approach, they can be referred to as "quasi-repeated" units.

As used herein **"random peptide library"** refers to a set of polynucleotide sequences that encodes a set of random peptides, and to the set of random peptides encoded by those polynucleotide sequences, as well as the fusion proteins contain those random peptides.

As used herein, **"random peptide sequence"** refers to an amino acid sequence composed of two or more amino acid monomers and constructed by a stochastic or random process. A random peptide can include framework or scaffolding motifs, which may comprise invariant sequences.

As used herein, **"receptor"** refers to a molecule that has an affinity for a given ligand. Receptors can be naturally occurring or synthetic molecules. Receptors can be employed in an unaltered state or as aggregates with other species. Receptors can be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of receptors include, but are not limited to, antibodies, including monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells, or other materials), cell membrane receptors, complex carbohydrates and glycoproteins, enzymes, and hormone receptors.

"Recombinant" enzymes refer to enzymes produced by recombinant DNA techniques, i.e., produced from cells transformed by an exogenous DNA construct encoding the desired enzyme. **"Synthetic"** enzymes are those prepared by chemical synthesis.

The term **"related polynucleotides"** means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

"Reductive reassortment", as used herein, refers to the increase in molecular diversity that is accrued through deletion (and/or insertion) events that are mediated by repeated sequences.

The following terms are used to describe the sequence relationships between two or more polynucleotides: **"reference sequence," "comparison window," "sequence identity," "percentage of sequence identity,"** and **"substantial identity."**

A **"reference sequence"** is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (i.e., a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a **"comparison window"** to identify and compare local regions of sequence similarity.

"Repetitive Index (RI)", as used herein, is the average number of copies of the quasi-repeated units contained in the cloning vector.

The term **"restriction site"** refers to a recognition sequence that is necessary for the manifestation of the action of a restriction enzyme, and includes a site of catalytic cleavage. It is appreciated that a site of cleavage may or may not be contained within a portion of a restriction site that comprises a low ambiguity sequence (i.e. a sequence containing the principal determinant of the frequency of occurrence of the restriction site). Thus, in many cases, relevant restriction sites contain only a low ambiguity sequence with an internal cleavage site (e.g. G/AATTC in the EcoR I site) or an immediately adjacent cleavage site (e.g. /CCWGG in the EcoR II site). In other cases, relevant restriction enzymes [e.g. the Eco57 I site or CTGAAG(16/14)] contain a low ambiguity sequence

(e.g. the CTGAAG sequence in the Eco57 I site) with an external cleavage site (e.g. in the N₁₆ portion of the Eco57 I site). When an enzyme (e.g. a restriction enzyme) is said to "cleave" a polynucleotide, it is understood to mean that the restriction enzyme catalyzes or facilitates a cleavage of a polynucleotide.

In a non-limiting aspect, a "selectable polynucleotide" is comprised of a 5' terminal region (or end region), an intermediate region (i.e. an internal or central region), and a 3' terminal region (or end region). As used in this aspect, a 5' terminal region is a region that is located towards a 5' polynucleotide terminus (or a 5' polynucleotide end); thus it is either partially or entirely in a 5' half of a polynucleotide. Likewise, a 3' terminal region is a region that is located towards a 3' polynucleotide terminus (or a 3' polynucleotide end); thus it is either partially or entirely in a 3' half of a polynucleotide. As used in this non-limiting exemplification, there may be sequence overlap between any two regions or even among all three regions.

The term "sequence identity" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. This "substantial identity", as used herein, denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence having at least 80 percent sequence identity, preferably at least 85 percent identity, often 90 to 95 percent sequence identity, and most commonly at least 99 percent sequence identity as compared to a reference sequence of a comparison window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

As known in the art "similarity" between two enzymes is determined by comparing the amino acid sequence and its conserved amino acid substitutes of one enzyme to the sequence of a second enzyme. Similarity may be determined by procedures which are well-known in the art, for example, a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information).

As used herein, the term "single-chain antibody" refers to a polypeptide comprising a V_H domain and a V_L domain in polypeptide linkage, generally linked via a spacer peptide (e.g., [Gly-Gly-Gly-Gly-Ser]_x), and which may comprise additional amino acid sequences at the amino- and/or carboxy- termini. For example, a single-chain antibody may comprise a tether segment for linking to the encoding polynucleotide. As an example, a scFv is a single-chain antibody. Single-chain antibodies are generally proteins consisting of one or more polypeptide segments of at least 10 contiguous amino substantially encoded by genes of the immunoglobulin superfamily (e.g., see Williams and Barclay, 1989, pp. 361-368, which is incorporated herein by reference), most frequently encoded by a rodent, non-human primate, avian, porcine bovine, ovine, goat, or human heavy chain or light chain gene sequence. A functional single-chain antibody generally contains a sufficient portion of an immunoglobulin superfamily gene product so as to retain the property of binding to a specific target molecule, typically a receptor or antigen (epitope).

The members of a pair of molecules (e.g., an antibody-antigen pair or a nucleic acid pair) are said to "specifically bind" to each other if they bind to each other with greater affinity than to other, non-specific molecules. For example, an antibody raised against an antigen to which it binds more efficiently than to a non-specific protein can be described as specifically binding to the antigen. (Similarly, a nucleic acid probe can be described as specifically binding to a nucleic acid target if it forms a specific duplex with the target by base pairing interactions (see above).)

"Specific hybridization" is defined herein as the formation of hybrids between a first polynucleotide and a second polynucleotide (e.g., a polynucleotide having a distinct but substantially identical sequence to the first polynucleotide), wherein substantially unrelated polynucleotide sequences do not form hybrids in the mixture.

The term **"specific polynucleotide"** means a polynucleotide having certain end points and having a certain nucleic acid sequence. Two polynucleotides wherein one polynucleotide has the identical sequence as a portion of the second polynucleotide but different ends comprises two different specific polynucleotides.

"Stringent hybridization conditions" means hybridization will occur only if there is at least 90% identity, preferably at least 95% identity and most preferably at least 97% identity between the sequences. *See Sambrook et al, 1989, which is hereby incorporated by reference in its entirety.*

Also included in the invention are polypeptides having sequences that are **"substantially identical"** to the sequence of a phytase polypeptide, such as one of SEQ ID 1. A **"substantially identical"** amino acid sequence is a sequence that differs from a reference sequence only by conservative amino acid substitutions, for example, substitutions of one amino acid for another of the same class (*e.g.*, substitution of one hydrophobic amino acid, such as isoleucine, valine, leucine, or methionine, for another, or substitution of one polar amino acid for another, such as substitution of arginine for lysine, glutamic acid for aspartic acid, or glutamine for asparagine).

Additionally a **"substantially identical"** amino acid sequence is a sequence that differs from a reference sequence or by one or more non-conservative substitutions, deletions, or insertions, particularly when such a substitution occurs at a site that is not the active site the molecule, and provided that the polypeptide essentially retains its behavioural properties. For example, one or more amino acids can be deleted from a phytase polypeptide, resulting in modification of the structure of the polypeptide, without significantly altering its biological activity. For example, amino- or carboxyl-terminal amino acids that are not required for phytase biological activity can be removed. Such modifications can result in the development of smaller active phytase polypeptides.

The present invention provides a **"substantially pure enzyme"**. The term **"substantially pure enzyme"** is used herein to describe a molecule, such as a polypeptide (*e.g.*, a phytase polypeptide, or a fragment thereof) that is substantially free of other

proteins, lipids, carbohydrates, nucleic acids, and other biological materials with which it is naturally associated. For example, a substantially pure molecule, such as a polypeptide, can be at least 60%, by dry weight, the molecule of interest. The purity of the polypeptides can be determined using standard methods including, e.g., polyacrylamide gel electrophoresis (e.g., SDS-PAGE), column chromatography (e.g., high performance liquid chromatography (HPLC)), and amino-terminal amino acid sequence analysis.

As used herein, "**substantially pure**" means an object species is the predominant species present (i.e., on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

As used herein, the term "**variable segment**" refers to a portion of a nascent peptide which comprises a random, pseudorandom, or defined kernal sequence. A variable segment" refers to a portion of a nascent peptide which comprises a random pseudorandom, or defined kernal sequence. A variable segment can comprise both variant and invariant residue positions, and the degree of residue variation at a variant residue position may be limited: both options are selected at the discretion of the practitioner. Typically, variable segments are about 5 to 20 amino acid residues in length (e.g., 8 to 10), although variable segments may be longer and may comprise antibody portions or receptor proteins, such as an antibody fragment, a nucleic acid binding protein, a receptor protein, and the like.

The term "**wild-type**" means that the polynucleotide does not comprise any mutations. A "wild type" protein means that the protein will be active at a level of activity found in nature and will comprise the amino acid sequence found in nature.

The term “**working**”, as in “**working sample**”, for example, is simply a sample with which one is working. Likewise, a “**working molecule**”, for example is a molecule with which one is working.

F. DETAILED DESCRIPTION OF THE INVENTION

1. GENOMIC CHARACTERIZATION METHODS

In one aspect, this invention describes a new method to sequence DNA. The improvements over the existing DNA sequencing technologies are high speed, high throughput, no electrophoresis and gel reading artifacts due to the complete absence of an electrophoretic step, and no costly reagents involving various substitutions with stable isotopes. The invention utilizes the Sanger sequencing strategy and assembles the sequence information by analysis of the nested fragments obtained by basespecific chain termination via their different molecular masses using mass spectrometry, as for example, MALDI or ES mass spectrometry. A further increase in throughput can be obtained by introducing massmodifications in the oligonucleotide primer, chain-terminating nucleoside triphosphates and/or in the chainelongating nucleoside triphosphates, as well as using integrated tag sequences which allow multiplexing by hybridization of tag specific probes with mass differentiated molecular weights.

The present invention pertains to a method for sequencing genomes. The method comprises the steps of obtaining nucleic acid material from a genome. Then there is the step of constructing a clone library and one or more probe libraries from the nucleic acid material. Next there is the step of comparing the libraries to form comparisons. Then there is the step of combining the comparisons to construct a map of the clones relative to the genome. Next there is the step of determining the sequence of the genome by means of the map.

The present invention also pertains to a system for sequencing a genome. The system comprises a mechanism for obtaining nucleic acid material from a genome. The system also comprises a mechanism for constructing a clone library and one or more probe libraries. The constructing mechanism is in communication with the nucleic acid material from a genome. Additionally, the system comprises a mechanism for comparing said libraries to form comparisons. The comparing mechanism is in communication with the said libraries. The system also comprises a mechanism for combining the comparisons to construct a map of the clones relative to the genome. The said combining mechanism is in communication with the comparisons. Further, the system comprises a mechanism for determining the

sequence of the genome by means of said map. The said determining mechanism is in communication with said map. The present invention additionally pertains to a method for producing a gene of a genome.

An efficient method for sequencing large fragments of DNA is described. A subclone path through the fragment is first identified; the collection of subclones that define this path is then sequenced using transposon-mediated direct sequencing techniques to an extent sufficient to provide the complete sequence of the fragment.

Improved techniques are provided for DNA sequencing, and particularly for sequencing of the entire human genome. Different base-specific reactions are utilized to use different sets of DNA fragments from a piece of DNA of unknown sequence. Each of the different sets of DNA fragments has a common origin and terminates at a particular base along the unknown sequence. The molecular weight of the DNA fragments in each of the different sets is detected by a matrix assisted laser absorption mass spectrometer to determine the sequence of the different bases in the DNA. The methods and apparatus of the present invention provide a relatively simple and low cost technique which may be automated to sequence thousands of gene bases per hour, and eliminates the tedious and time consuming gel electrophoresis separation technique conventionally used to determine the masses of DNA fragments.

Processes and kits for simultaneously amplifying and sequencing nucleic acid molecules, and performing high throughput DNA sequencing are described.

A new contiguous genome sequencing method is described which allows the contiguous sequencing of a very long DNA without need to be subcloned. It uses the basic PCR technique but circumvents the usual need of this technique for the knowledge two primers for contiguous sequencing, enabling the knowledge of only one primer sufficient. The present invention makes it possible to PCR amplify a DNA adjacent to a known sequence with which one primer can be made without the knowledge of the second primer binding site present in the unknown sequence. The present invention could thus be used to contiguously sequence a very long DNA such as that contained in a YAC clone or a cosmid clone, without the need for subcloning smaller fragments, using the standard PCR technique. It can also be used to sequence a whole chromosome or genome without any need to subclone it.

Methods and means are provided for the massively parallel characterization of complex molecules and of molecular recognition phenomena with parallelism and

redundancy attained through single molecule examination methods. Applications include ultra-rapid genome sequencing, affinity characterization, pathogen characterization and detection means for clinical use and use in the development and construction of cybernetic immune systems. Novel methods for single molecule examination and manipulation are provided, including scanned beam light microscopic means and methods, and detection means availing of optoelectronic array devices. Various apparatus for rate control, including stepping control for various reactions are combined with molecular recognition, signal amplification and single molecule examination methods. Inclusion of internal control in samples, algorithm-based dynamically responsive manipulation controls, and sample redundancy, are availed to provide an arbitrarily high degree of accuracy in final data.

1.1 SEQUENCING

The present invention relates to sequencing of DNA and is in the field of determining the nucleotide sequence of large segments of DNA. More specifically, the invention provides an improved method to obtain the complete nucleotide sequence of genomic DNA provided in fragments of over 30 kb.

The present invention pertains to a process for determining the DNA sequence of the genome of an organism. And more particularly, the invention relates to the sequencing of the entire human genome.

More specifically, the present invention is related to constructing clone maps of organisms, and then using these maps to direct the sequencing effort. The invention also pertains to systems that can effectively use this sequence and map information.

The invention relates to the massively parallel single molecule examination of associations or reactions between large numbers of first complex molecules, which may be diverse, and second single or plural probing molecules, which may or may not be diverse, with applications to biology, biotechnology, pharmacology, immunology, the novel field of cybernetic immunology, molecular evolution, cybernetic molecular evolution, genomics, comparative genomics, enzymology, clinical enzymology, pathology, medical research, and clinical medicine.

The present invention has applications in the area of polynucleotide sequence determination, including DNA sequencing.

1.2 SEQUENCING METHODS

1.2.1 Importance of DNA sequencing:

Current knowledge regarding gene structure, the control of gene activity and the function of cells on a molecular level all arose based on the determination of the base sequence of millions of DNA molecules. DNA sequencing is still critically important in research and for genetic therapies and diagnostics, (e.g., to verify recombinant clones and mutations).

DNA, a polymer of deoxyribonucleotides, is found in all living cells and some viruses. DNA is the carrier of genetic information, which is passed from one generation to the next by homologous replication of the DNA molecule. Information for the synthesis of all proteins is encoded in the sequence of bases in the DNA. DNA sequence information represents the information required for gene organization and regulation of most life forms. Accordingly, the development of reliable methodology for sequencing DNA has contributed significantly to an understanding of gene structure and function.

Since the genetic information is represented by the sequence of the four DNA building blocks deoxyadenosine- (dpA), deoxyguanosine- (dpG), deoxycytidine- (dpC) and deoxythymidine-5'-phosphate (dpT), DNA sequencing is one of the most fundamental technologies in molecular biology and the life sciences in general. The ease and the rate by which DNA sequences can be obtained greatly affects related technologies such as development and production of new therapeutic agents and new and useful varieties of plants and microorganisms via recombinant DNA technology. In particular, unraveling the DNA sequence helps in understanding human pathological conditions including genetic disorders, cancer and AIDS. In some cases, very subtle differences such as a one nucleotide deletion, addition or substitution can create serious, in some cases even fatal., consequences. Recently, DNA sequencing has become the core technology of the Human Genome Sequencing Project (e.g., J.E. Bishop and M. Waldholz, 1991, *Genome: The Story of the Most Astonishing Scientific Adventure of Our Time - The Attempt to Map All the Genes in the Human Body*, Simon & Schuster, New York). Knowledge of the complete human genome DNA sequence will certainly help to understand, to diagnose, to prevent and to treat human diseases. To be able to tackle successfully the determination of the approximately 3 billion base pairs of the human genome in a reasonable time frame

and in an economical way, rapid, reliable, sensitive and inexpensive methods need to be developed, which also offer the possibility of automation. The present invention provides such a technology. The need for highly rapid, accurate, and inexpensive sequencing technology is nowhere more apparent than in a demanding sequencing project such as the human genome project.

The present invention relates to the field of nucleic acid analysis, detection, and sequencing. More specifically, in one embodiment the invention provides improved techniques for synthesizing arrays of nucleic acids, hybridizing nucleic acids, detecting mismatches in a double-stranded nucleic acid composed of a single-stranded probe and a target nucleic acid, and determining the sequence of DNA or RNA or other polymers.

A human being has 23 pairs of chromosomes consisting of a total of about 100,000 genes. The human genome consists of those genes. A single gene which is defective may cause an inheritable disease, such as Huntington's disease, Tay-Sachs disease or cystic fibrosis. The human chromosomes consist of large organic linear molecules of double-strand DNA (deoxyribonucleic acid) with a total length of about 3.3 billion "base pairs". The base pairs are the chemicals that encode information along DNA. A typical gene may have about 30,000 base pairs. By correlating the inheritance of a "marker" (a distinctive segment of DNA) with the inheritance of a disease, one can find a mutant (abnormal) gene to within one or two million base pairs. This opens the way to clone the DNA segment, test its activity, follow its inheritance, and diagnose carriers and future disease victims.

The mapping of the human genome is to accurately determine the location and composition of each of the 3.3 billion bases. The complexity and large scale of such a mapping has placed it, in terms of cost, effort and scientific potential of such projects, as one of the largest and most important projects of the 1990's and beyond.

Recent reviews of today's methods together with future directions and trends are given by Barrell (The FASEB Journal 1, 40-45 (1991)), and Trainor (Anal. Chem. 62, 418-26 (1990)).

1.2.2 Previously developed methods:

The problem of DNA sequence analysis is that of determining the order of the four bases on the DNA strands. DNA sequencing is a technique by which the four DNA nucleotides (characters) in a linear DNA sequence is ordered by chemical and biochemical means. Generally, strategies for determining the nucleotide sequence of DNA involve the generation of a DNA substrate i.e., DNA fragments suitable for sequencing a region of the DNA, enzymatic or chemical reactions, and analysis of DNA fragments that have been separated according to their lengths to yield sequence information. More specifically, to sequence a given region of DNA, labeled DNA fragments are typically generated in four separate reactions. In each of the four reactions, the DNA fragments typically have one fixed end and one end that terminates sequentially at each of the four nucleotide bases, respectively. The products of each reaction are fractionated by gel electrophoresis on adjacent lanes of a polyacrylamide gel. As all of the nucleotides are represented among the four lanes, the sequence of a given region of DNA can be determined from the four "ladders" of DNA fragments. The present status of techniques for determining such sequences is described in some detail in an article by Lloyd M. Smith published in the American Biotechnology Laboratory, Volume 7, Number 5, May 1989, pp 10-17. Since the early 1970's, two methods have been developed for the determination of DNA sequence: (1) the enzymatic chain-termination sequencing method, which relies on the template directed incorporation of nucleotides which themselves do not supply the necessary chemical functionalities required for subsequent enzymatic polymerization of a daughter strand polynucleotide, developed by Sanger and colleagues (F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors." *Proc. Natl. Acad. Sci. USA*, 74:5463-5467 (1977)), which is most commonly used for sequence determination; and (2) the base-specific chemical degradation (modification and cleavage) method, developed by Maxam and Gilbert (A. M. Maxam, and W. Gilbert, "A new method of sequencing DNA." *Proceedings of the National Academy of Sciences, USA*, 74:560-564 (1977)), which similarly yields polynucleotide molecules terminated at sites containing a specific base according to the chemical treatment applied to the sample. Both of these techniques are based on similar principals, and employ gel electrophoresis to separate DNA fragments of different lengths with high resolution. On these gels it is thus possible to separate a DNA

fragment 600 bases in length from one 601 bases in length. No distinct method preferable to these has yet been validated. Both methods require a large number of complex manipulations, such as isolation of homogeneous DNA fragments, elaborate and tedious preparation of samples, preparation of a separating gel, application of samples to the gel, electrophoresing the samples on the gel, working up of the finished gel, and analysis of the results of the procedure.

1. 2.2.1 Chemical/Maxam and Gilbert method for sequencing:

In the chemical method, the DNA strand is isotropically labeled on one end, broken down into smaller fragments at sequence locations ending with a particular nucleotide (A, T, C, or G) by chemical means, and the fragments ordered based on this information. Base specific modifications result in a base specific cleavage of the radioactive or fluorescently labeled DNA fragment. After the DNA substrate is end labeled, it is subjected to chemical reactions designed to cleave the DNA at positions adjacent to a given base or bases. The labeled DNA fragments will, therefore, have a common labeled terminus while the unlabeled termini will be defined by the positions of chemical cleavage. This results in the generation of DNA fragments (four sets of nested fragments) which can be separated according to length by polyacrylamide gel electrophoresis (PAGE) and identified. Alternatively, unlabeled DNA fragments can be separated after complete restriction digestion and partial chemical cleavage of the DNA, and hybridized with probes homologous to a region near the region of the DNA to be sequenced. See, Church et al., Proc. Natl. Acad. Sci., 81:1991 (1984). After autoradiography, the sequence can be read directly since each band (fragment) in the gel originates from a base specific cleavage event. Thus, the fragment lengths in the four "ladders" directly translate into a specific position in the DNA sequence.

1. 2.2.2 Enzymatic/Sanger method for sequencing:

In the enzymatic method, the four base specific sets of DNA fragments are formed by starting with a primer/template system elongating the primer into the unknown DNA sequence area and thereby copying the template and synthesizing complementary strands using a DNA polymerase in the presence of chain-terminating reagents. The chain-terminating event is achieved by incorporating into the four separate reaction mixtures in addition to the four normal deoxynucleoside triphosphates, dATP, dGTP, dTTP and dCTP, only one of the chain-terminating dideoxynucleoside triphosphates, ddATP, ddGTP, ddTTP or ddCTP, respectively, in a

limiting small concentration. The incorporation of a ddNTP lacking the 3' hydroxyl function into the growing DNA strand by the enzyme DNA polymerase leads to chain termination through preventing the formation of a 3'-5'-phosphodiester bond by DNA polymerase. Due to the random incorporation of the ddNTPs, each reaction leads to a population of base specific terminated fragments of different lengths, which all together represent the sequenced DNA-molecule. The four sets of resulting fragments produce, after electrophoresis, four base specific ladders from which the DNA sequence can be determined.

In the enzymatic method, the following basic steps are involved:

(i) annealing an oligonucleotide primer to a suitable single or denatured double stranded DNA template; (ii) extending the primer with DNA polymerase in four separate reactions, each containing one - labeled dNTP or ddNTP (alternatively a labeled primer can be used), a mixture of unlabeled dNTPs, and one chain-terminating dideoxynucleoside- 5'-triphosphate (ddNTP); (iii) resolving the four sets of reaction products, which include a distribution of DNA fragments having primer-defined 5' termini and differing dideoxynucleotides at the 3' termini, on a high resolution polyacrylamide-urea gel; and (iv) producing an auto radiographic image of the gel that can be examined to infer the DNA sequence. Alternatively, fluorescently labeled primers or nucleotides can be used to identify the reaction products. Known dideoxy sequencing methods utilize a DNA polymerase such as the Klenow fragment of *E. coli* DNA polymerase, a DNA polymerase from *Thermus aquaticus*, reverse transcriptase, a modified T7 DNA polymerase, or the Taq polymerase.

1.2.2.3 Similarities, differences and other details of the two methods:

The two sequencing methods differ in the techniques employed to produce the DNA fragments, but are otherwise similar. In the Maxam-Gilbert method, four different base-specific reactions are performed on portions of the DNA molecules to be sequenced, to produce four sets of radiolabeled DNA fragments. These four fragment sets are each loaded in adjacent lanes of a polyacrylamide slab gel, and are separated by electrophoresis. Autoradiographic imaging of the pattern of the radiolabeled DNA bands in the gel reveals the relative size, corresponding to band mobilities, of the fragments in each lane, and the DNA sequence is deduced from this pattern.

While numerous modifications and improvements to the strategies referred to above have been developed, most sequencing techniques require the presence of a known primer binding site for every 300 to 500 nucleotides to be sequenced either, for example, for initiation of DNA synthesis or for hybridization to different length DNA fragments having a common end. However, as such approaches utilize a "ladder" of DNA fragments containing the primer binding site (or its complement), the amount of sequence information that can be obtained is limited by the present inability to resolve DNA fragments greater than 500 nucleotides in length on sequencing gels.

Both of these methods yield a population of molecules comprising a nested set which together may be analyzed to determine the base sequence of the sample. At least one of these two techniques is employed in essentially every laboratory concerned with molecular biology, and together they have been employed to sequence more than 26 million bases of DNA. Currently a skilled biologist can produce about 30,000 bases of finished DNA sequence per year under ideal conditions.

These methods and several variations thereupon, as well as their severe limitations with respect to the economy and rapidity of accumulation of sequence data, are well known to those in the relevant arts. Various lower resolution techniques, generally falling within the category termed genome mapping, have been developed to circumvent these limitations for applications where more "broad spectrum" examination of genetic material is required but less detailed information about sequence will suffice.

1. 2.2.4 Cloning/Subcloning steps:

On the upfront end, the DNA to be sequenced has to be fragmented into sequencable pieces of currently not more than 500 to 1000 nucleotides. Starting from a genome, this is a multi-step process involving cloning and subcloning steps using different and appropriate cloning vectors such as YAC, cosmids, plasmids and M13 vectors (Sambrook et al., *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 1989). Finally, for Sanger sequencing, the fragments of about 500 to 1000 base pairs are integrated into a specific restriction site of the replicative form I (RF I) of a derivative of the M13 bacteriophage (Vieria and Messing, *Gene* 19, 259(1982)) and then the double-stranded form is transformed to the single-stranded circular form to serve as a template for the Sanger sequencing

process having a binding site for a universal primer obtained by chemical DNA synthesis (Sinha, Biernat, McManus and Köster, *Nucleic Acids Res.* 12, 4539-57 (1984); U.S. Patent No. 4725677 upstream of the restriction site into which the unknown DNA fragment has been inserted. Under specific conditions, unknown DNA sequences integrated into supercoiled double-stranded plasmid DNA can be sequenced directly by the Sanger method (Chen and Seeburg, *DNA* 4, 165-170 (1985)) and Lim et al., *Gene Anal., Techn.* 5, 32-39 (1988), and, with the Polymerase Chain Reaction (PCR) (PCR Protocols- A Guide to Methods and Applications. Innis et al., editors, Academic Press, San Diego (1990)) cloning or subcloning steps could be omitted by directly sequencing off chromosomal DNA by first amplifying the DNA segment by PCR and then applying the Sanger sequencing method (Innis et al., *Proc. Nat. Acad. Sci. USA* 85, 9436-9440 (1988)). In this case, however, the DNA sequence in the interested region must be known at least to the extent to bind a sequencing primer.

1. 2.2.5 Methodology described by Guo and Wu

Methodology described by Guo and Wu, *Nucleic Acids Res.*, 10:2065 (1982); and *Meth. Enz.*, 100:60 (1983), which is not dependent upon primer binding sites, is highly desirable for sequencing DNA greater than 500 nucleotides. This method involves partially digesting linear double stranded DNA with *E. coli* exonuclease III to produce DNA fragments with 3' ends shortened to varying lengths, performing the dideoxy primer extension reactions of Sanger, *supra*, with the shortened 3' ends as primers for DNA synthesis, and digesting the DNA with a selected restriction enzyme that cleaves near one end of the molecule adjacent to, but not within, the labeled region of DNA. By digestion of the DNA with a selected restriction enzyme, the labeled DNA strands from one end of the molecule are made small enough to be resolved on a sequencing gel. Each successive deletion in length, therefore, brings "new" regions of the target DNA into sequencing range.

However, certain disadvantages inherent in the methodology of Guo and Wu, *supra*, limit its usefulness for the large scale sequencing of DNA. For example, this approach depends upon the selection of appropriate restriction enzymes which cleave at restriction sites in close proximity to particular *E. coli* exonuclease III endpoints, but not within the labeled DNA as this would result in two or more superimposed sequence ladders. The selection of appropriate restriction enzymes generally requires,

therefore, the restriction mapping of DNA fragments to identify sites in close proximity to the numerous exonuclease III endpoints. However, the determination of restriction maps tends to be both time consuming and labor intensive. Specifically, restriction mapping to the resolution needed for DNA sequencing involves the digestion of each region of DNA with combinations of 20 or more enzymes to uncover the relative position of restriction sites. This may require over 100 enzymatic reactions followed by numerous electrophoretic separations. Further, significant amounts of DNA are consumed in the mapping process and interpretation of the data generally requires a substantial amount of time.

1. 2.3 3'-hydroxy-protected and labeled nucleotides:

A modified nucleotide compound possessing two properties particularly useful for purposes of the present invention has been described by N. Williams and P.S. Coleman. This compound is 3'-O-(4-benzoyl)benzoyl adenosine 5'-triphosphate. This nucleotide bears a 3' protecting group linked via an ester function which should be susceptible to hydrolysis by appropriate chemical treatments. The protecting moiety is suitable for photoactivation, and this property was utilized by those investigators to probe the structure of mitochondrial F_1 -ATPase, indicating that this analog will interact properly with at least some enzymes. Under appropriate circumstances, the protecting moiety may also serve as a label.

Very recently, B Canard and R.S. Sarfati have described similar nucleotides, here comprising all four nucleobases, with chemically removable 3'-hydroxyl protecting groups. Said protecting groups comprise various fluorescent dye moieties. These investigators have shown that these compounds may be added to appropriately primed polynucleotides by polymerases according to Watson-Crick base-pairing rules, and serve to terminate chain elongation in a manner which may be reversed by removal of said protecting groups by appropriate chemical treatments, admitting resumption of polymerization. These workers propose that such compounds may form the basis of a novel sequencing methodology availing stepping control by means of said removable protecting groups and detection of labels following their release from the nascent strand by appropriate chemical treatment. Such a method, while a potential advance over electrophoretic resolution methods, does not avail of great parallelism because only one molecule or an identical population of molecules may be sequenced at once (within a single vessel) by such a method, due to the release of the labeling moiety prior to detection, according to this proposed scheme. Further, this limitation requires that any attempt to avail of parallelism entail elaborate parallel fluidics. Low or no parallelism entails that stepping rate will be critical to the throughput attained with such a sequencing scheme. The results published by these authors suggests that the rate of chemical removal of 3'-hydroxy protecting groups (less than 90% removal after 10 minutes of treatment with 0.1M NaOH) will be unacceptably low for such an inherently serial sequencing scheme.

Additional references regarding such compounds and in most instances their properties as substrates for various enzymes including polymerases have been found

in the biological literature: Churchich, J.E.; 1995. *Eur. J. Biochem.*, 231:736. Metzket, M.L.; Gibbs, R.A.; et al.; 1994. *Nucleic Acids Research*, 22:4259. Beabealashvili, R.S.; Kukhanova, M.K.; et al.; 1986. *Biochimica et Biophysica Acta*, 868:136. Chidgeavadze, Z.G.; Kukhanova, M.K.; et al.; 1986. *Biochimica et Biophysica Acta*, 868:145. Hiratsuka, T; 1983. *Biochimica et Biophysica Acta*, 742:496. Jeng, S.J.; Guillory, R.J.; 1975. *J. Supramolecular Structure*, 3:448.

1. 2.4 Related Base Addition Sequencing Schemes:

Various other investigators have also independently devised polynucleotide sequencing methodologies which depend on the addition of a polymerization terminating labeled nucleotide to a primed or elongated daughter strand on a polynucleotide sample with template dependent polynucleotide polymerases. Most, but not all, of these methods (referred to herein as previously disclosed base-addition sequencing schemes) avail nucleotide triphosphate monomers with some base-specific label which may be removed by some deprotection treatment. It must be emphasized that all of these other previously disclosed base-addition sequencing schemes examine not single molecules individually but rather large homogeneous populations of substantially identical molecules, wherein the observed signal used to identify label type originates from the totality of such a population of molecules rather than an individual molecule. It must be further emphasized that conventional usage does not generally reveal this distinction: phrases such as "a molecule" or "a sample molecule" refer not to an individual molecule considered separately or in isolation from other molecules including separately from other molecules of identical composition and structure, but to populations comprising millions or more molecules of identical structure. A careful reading of these prior disclosures reveals that these investigators are not working with samples consisting of single molecules but rather with samples comprising a plurality of identical molecules. In particular, even where these investigators do not (as is consistent with conventional usage) explicitly note this point, they take measures which would apply only to samples of pluralities of identical molecules, and do not take measures associated with working with single molecules.

1. 2.5 Labeling

1. 2.5.1 Sequencing from PAGE using radioisotopes:

In order to be able to read the sequence from PAGE, detectable labels have to be used in either the primer (very often at the 5'-end) or in one of the deoxynucleoside triphosphates, dNTP. Using radioisotopes such as ^{32}P , ^{33}P , or ^{35}S is still the most frequently used technique. After PAGE, the gels are exposed to X-ray films and silver grain exposure is analyzed. The use of radioisotopic labeling creates several problems. Most labels useful for autoradiographic detection of sequencing fragments have relatively short half-lives which can limit the useful time of the labels. The emission high energy beta radiation, particularly from ^{32}P , can lead to breakdown of the products via radiolysis so that the sample should be used very quickly after labeling. In addition, high energy radiation can also cause a deterioration of band sharpness by scattering. Some of these problems can be reduced by using the less energetic isotopes such as ^{33}P or ^{35}S (see, e.g., Ornstein et al., *Biotechniques* 2, 476 (1985)). Here, however, longer exposure times have to be tolerated. Above all, the use of radioisotopes poses significant health risks to the experimentalist and, in heavy sequencing projects, decontamination and handling the radioactive waste are other severe problems and burdens.

1. 2.5.2 Integration of non-radioactive labeling techniques into partly automated DNA sequencing:

In response to the above mentioned problems related to the use of radioactive labels, non-radioactive labeling techniques have been explored and, in recent years, integrated into partly automated DNA sequencing procedures. All these improvements utilize the Sanger sequencing strategy. The fluorescent label can be tagged to the primer (Smith et al., *Nature* M, 674-679 (1986) and EPO Patent No. 873 00998.9; Du Pont De Nemours EPO Application No. 03 59225; Ansorge et al., *L Biochem. Biophys. Method* 13, 325-32 (1986)) or to the chain-terminating dideoxynucleoside triphosphates (Prober et al. *Science* M, 336-41 (1987); Applied Biosystems, PCT Application WO 91/05060). Based on either labeling the primer or the ddNTP, systems have been developed by Applied Biosystems (Smith et al., *Science* 235, G89 (1987); U.S. Patent Nos. 570973 and 689013), Du Pont De Nemours (Prober et al., *Science* 238, 336-341 (1987); U.S. Patents Nos. 881372 and 57566), Pharmacia-LKB (Ansorge et al. *Nucleic Acids Res.* 15-, 4593-4602 (1987)

and EMBL Patent Application DE P3724442 and P3805808.1) and Hitachi (JP 1-90844 and DE 4011991 A1). A somewhat similar approach was developed by Brumbaugh et al. (Proc. Natl. Sci. USA 85, 5610-14 (1988) and U.S. Patent No. 4,729,947). An improved method for the Du Pont system using two electrophoretic lanes with two different specific labels per lane is described (PCT Application W092/02635). A different approach uses fluorescently labeled avidin and biotin labeled primers. Here, the sequencing ladders ending with biotin are reacted during electrophoresis with the labeled avidin which results in the detection of the individual sequencing bands (Brumbaugh et al., U.S. Patent No. 594676).

More recently even more sensitive non-radioactive labeling techniques for DNA using chemiluminescence triggerable and amplifiable by enzymes have been developed (Beck, O'Keefe, Coull and Köster, Nucleic Acids Res. 7, 5115- 5123 (1989) .L7 and Beck and Köster, Anal. Chem. 62 2258-2270 (1990)). These labeling methods were combined with multiplex DNA sequencing (Church et al., Science 240, 185-188 (1988) to provide for a strategy aimed at high throughput DNA sequencing (Köster et al., Nucleic Acids Res. Symposium Ser. No. 24, 318-321 (1991), University of Utah, PCT Application No. WO 90/15883); this strategy still suffers from the disadvantage of being very laborious and difficult to automate.

1. 2.5.2.1 Fluorescent labeling ing in methods for automated DNA sequencing

Of particular interest in DNA sequencing are methods of automated sequencing, in which fluorescent labels are employed to label the size separated fragments or primer extension products of the enzymatic method. Currently, three different methods are used for automated DNA sequencing. In the first method, the DNA fragments are labeled with one fluorophore and then run in adjacent sequencing lanes, one lane for each base. See Ansorge et al., Nucleic Acids Res. (1987)15:4593-4602. In the second methods, the DNA fragments are labeled with oligonucleotide primers tagged with four fluorophores and all of the fragments are run in one lane. See Smith et al., Nature (1986) 321:674- 679. In the third method, each of the different chain terminatina dideoxynucleotides is labeled with a different fluorophore and all of the fragments are run in one lane. See Prober et al., Science (1987) 238:336-341. The first method has the potential problems of lane-to-lane variations as well as a low throughput. The second and third methods require that the four dyes be well excited by one laser source, and that they have distinctly different emission

spectra. Otherwise, multiple lasers have to be used, increasing the complexity and the cost of the detection instrument.

With the development of Energy Transfer primers which offer strong fluorescent signals upon excitation at a common wavelength, the second method produces robust sequencing data in currently commercial available sequencers. However, even with the use of Energy Transfer primers, the second method is not entirely satisfactory. In the second method, all of the false terminated or false stop fragments are detected resulting in high backgrounds. Furthermore, with the second method it is difficult to obtain accurate sequences for DNA templates with long repetitive sequences. See Robbins et al., *Biotechniques* (1996) 20: 862-868.

The third method has the advantage of only detecting DNA fragments incorporated with a terminator. Therefore, backgrounds caused by the detection of false stops are not detected. However, the fluorescence signals offered by the dye-labeled terminators are not very bright and it is still tedious to completely clear up the excess of dye-terminators even with AmpliTaq DNA Polymerase (FS enzyme). Furthermore, non-sequencing fragments are detected, which contributes to background signal. Applied Biosystems Model 373 A DNA Sequencing System User Bulletin, November 17, P3, August 1990.

Thus, there is a need for the development of improved methodology which is capable of providing for highly accurate sequencing data, even for long repetitive sequences. Such methodology would ideally include a means for isolating the DNA sequencing fragments from the remaining components of the sequencing reaction mixtures such as salts, enzymes, excess primers, template and the like, as well as false stopped sequencing fragments and non-sequencing fragments resulting from contaminated RNA and nicked DNA templates.

1. 2.6 Simplifying DNA sequencing using solid supports:

In an attempt to simplify DNA sequencing, solid supports have been introduced. In most cases published so far, the template strand for sequencing (with or without PCR amplification) is immobilized on a solid support most frequently utilizing the strong biotin-avidin/streptavidin interaction (Orion-Yhtymä Oy, U.S. Patent No. 277643; M. Uhlen et al. Nucleic Acids Res. 16, 3025-38 (1988); Cemu Bioteknik, PCT Application No. WO 89/09282 and Medical Research Council, GB, PCT Application No. WO 92/03575). The primer extension products synthesized on the immobilized template strand are purified of enzymes, other sequencing reagents and by-products by a washing step and then released under denaturing conditions by loosing the hydrogen bonds between the Watson-Crick base pairs and subjected to PAGE separation. In a different approach, the primer extension products (not the template) from a DNA sequencing reaction are bound to a solid support via biotin/avidin (Du Pont De Nemours, PCT Application WO 91/11533). In contrast to the above mentioned methods, here, the interaction between biotin and avidin is overcome by employing denaturing conditions (formamide/EDTA) to release the primer extension products of the sequencing reaction from the solid support for PAGE separation. As solid supports, beads, (e. g., magnetic beads (Dynabeads) and Sepharose beads), filters, capillaries, plastic dipsticks (e.g., polystyrene strips) and microtiter wells are being proposed.

1. 2.7 Electrophoresis

1. 2.7.1 Drawbacks and limitations of polyacrylamide gel electrophoresis (PAGE):

All methods discussed so far have one central step in common: polyacrylamide gel electrophoresis (PAGE). In many instances, this represents a major drawback and limitation for each of these methods. Preparing a homogeneous gel by polymerization, loading of the samples, the electrophoresis itself, detection of the sequence pattern (e.g., by autoradiography), removing the gel and cleaning the glass plates to prepare another gel are very laborious and time-consuming procedures.

Moreover, the whole process is error-prone, difficult to automate, and, in order to improve reproducibility and reliability, highly trained and skilled personnel are required.

In the case of radioactive labeling, autoradiography itself can consume from hours to days. In the case of fluorescent labeling, at least the detection of the sequencing bands is being performed automatically when using the laser-scanning devices integrated into commercial available DNA sequencers. One problem related to the fluorescent labeling is the influence of the four different base-specific fluorescent tags on the mobility of the fragments during electrophoresis and a possible overlap in the spectral bandwidth of the four specific dyes reducing the discriminating power between neighboring bands, hence, increasing the probability of sequence ambiguities. Artifacts are also produced by base-specific interactions with the polyacrylamide gel matrix (Frank and Köster, *Nucleic Acids Res.* -6, 2069 (1979)) and by the formation of secondary structures which result in "band compressions" and hence do not allow one to read the sequence. This problem has, in part, been overcome by using 7-deazadeoxyguanosine triphosphates (Barr et al., *Biotechniques* 4, 428 (1986)). However, the reasons for some artifacts and conspicuous bands are still under investigation and need further improvement of the gel electrophoretic procedure.

1. 2.7.2 Capillary zone electrophoresis (CZE):

A recent innovation in electrophoresis is capillary zone electrophoresis (CZE) (Jorgenson et al., *J. Chromatography* 352, 337 (1986); Gesteland et al., *Nucleic Acids Res.* 18, 1415-1419 (1990)) which, compared to slab gel electrophoresis (PAGE), significantly increases the resolution of the separation, reduces the time for an

electrophoretic run and allows the analysis of very small samples. Here, however, other problems arise due to the miniaturization of the whole system such as wall effects and the necessity of highly sensitive on-line detection methods. Compared to PAGE, another drawback is created by the fact that CZE is only a "one-lane" process, whereas in PAGE samples in multiple lanes can be electrophoresed simultaneously.

1. 2.7.3 DNA sequencing without the electrophoretic step:

Analysis methods have heretofore relied on electrophoretic separation and resolution of the products of Sanger or Maxam and Gilbert reactions according to the length of said products. Analysis thus suffers all of the limitations associated with electrophoresis including limited separation range (i.e. limited dynamic range, where separative resolution is related exponentially to fractional differences in molecular length), limitations on parallelism, time requirements, etc., despite much effort in improving these separative methodologies. With presently available equipment and trained personnel, sequencing the human genome would require about 100 years of total effort if no other sequencing projects were done. While very useful, the present sequencing methods are extremely tedious and expensive, yet require the services of highly skilled scientists. Moreover, these methods utilize hazardous chemicals and radioactive isotopes, which have inhibited their consideration and further development. Large scale sequencing projects, as that of the human genome, thus appear to be impractical using these well-established techniques.

In addition to being slow, the present DNA sequencing techniques involve a large number of cumbersome handling steps which are difficult to automate. Recent improvements include replacing the radioactive labels with fluorescent tags. These developments have improved the speed of the process and have removed some of the tedious manual steps, although present technology continues to employ the relatively slow gel electrophoresis technique for separating the DNA fragments.

Due to the severe limitations and problems related to having PAGE as an integral and central part in the standard DNA sequencing protocol, several methods have been proposed to do DNA sequencing without an electrophoretic step. One approach calls for hybridization or fragmentation sequencing (Bains, *Biotechnology* 10, 757-58 (1992) and Mirzabekov et al., *FEBS Letters* 256, 118-122 (1989)) utilizing the specific hybridization of known short oligonucleotides (e.g., octadeoxynucleotides which gives 65,536 different sequences) to a complementary DNA sequence. Positive

hybridization reveals a short stretch of the unknown sequence. Repeating this process by performing hybridizations with all possible octadeoxynucleotides should theoretically determine the sequence. In a completely different approach, rapid sequencing of DNA is done by unilaterally degrading one single, immobilized DNA fragment by an exonuclease in a moving flow stream and detecting the cleaved nucleotides by their specific fluorescent tag via laser excitation (Jett et al., J. Biomolecular Structure & Dynamics 7, 301-309, (1989), United States Department of Energy, PCT Application No. WO 89/03432). In another system proposed by Hyman Anal. Biochem. 174, 423-436 (1988)), the pyrophosphate generated when the correct nucleotide is attached to the growing chain on a primer-template system is used to determine the DNA sequence. The enzymes used and the DNA are held in place by solid phases (DEAE-Sepharose and Sepharose) either by ionic interactions or by covalent attachment. In a continuous flow-through system, the amount of pyrophosphate is determined via bioluminescence (luciferase). A synthesis approach to DNA sequencing is also used by Tsien et al. (PCT Application No. WO 91/06678). Here, the incoming dNTP's are protected at the 3'-end by various blocking groups such as acetyl or phosphate groups and are removed before the next elongation step, which makes this process very slow compared to standard sequencing methods.

The template DNA is immobilized on a polymer support. To detect incorporation, a fluorescent or radioactive label is additionally incorporated into the modified dNTP's.

1.2.7.4 Apparatus to automate DNA sequencing without electrophoretic step(mass spectrometry):

PCT Application No. WO 91/06678 also describes an apparatus designed to automate the sequencing process.

Mass Spectrometry is a well known analytical technique which can provide fast and accurate molecular weight information on relatively complex mixtures of organic molecules. Mass spectrometry has historically had neither the sensitivity nor resolution to be useful for analyzing mixtures at high mass. A series of articles in 1988 by Hillenkamp and Karas do suggest that large organic molecules of about 10,000 to 100,000 Daltons may be analyzed in a time of flight mass spectrometer, although resolution at lower molecular weights is not as sharp as conventional

magnetic field mass spectrometry. Moreover, the Hillenkamp and Karas technique is very time-consuming, and requires complex and costly instrumentation.

Mass spectrometry, in general., provides a means of "weighing" individual molecules by ionizing the molecules in vacuo and making them "fly" by volatilization.

Under the influence of combinations of electric and magnetic fields, the ions follow trajectories depending on their individual mass (m) and charge (z). In the range of molecules with low molecular weight, mass spectrometry has long been part of the routine physical-organic repertoire for analysis and characterization of organic molecules by the determination of the mass of the parent molecular ion. In addition, by arranging collisions of this parent molecular ion with other particles (e.g., argon atoms), the molecular ion is fragmented forming secondary ions by the so-called collision induced dissociation (CID). The fragmentation pattern/pathway very often allows the derivation of detailed structural information. Many applications of mass spectrometric methods are known in the art, particularly in biosciences, and can be found summarized in *Methods in Enzymology*, Vol. 193: "Mass Spectrometry" Q.A. McCloskey, editor), 1990, Academic Press, New York.

Due to the apparent analytical advantages of mass spectrometry in providing high detection sensitivity, accuracy of mass measurements, detailed structural information by CID in conjunction with an MS/MS configuration and speed, as well as on-line data transfer to a computer, there has been considerable interest in the use of mass spectrometry for the structural analysis of nucleic acids. Recent reviews summarizing this field include K. H. Schram, "Mass Spectrometry of Nucleic Acid Components, Biomedical Applications of Mass Spectrometry" 34, 203-287 (1990); and P.F. Crain, "Mass Spectrometric Techniques in Nucleic Acid Research," *Mass Spectrometry Reviews* 9, 505-554 (1990). The biggest hurdle to applying mass spectrometry to nucleic acids is the difficulty of volatilizing these very polar biopolymers.

1. 2.8 Mass Spectrometry

1. 2.8.1 Limitation in applying mass spectrometry due to the difficulty of volatilizing nucleic acids:

Therefore, "sequencing" has been limited to low molecular weight synthetic oligonucleotides by determining the mass of the parent molecular ion and through this, confirming the already known sequence, or alternatively, confirming the known sequence through the generation of secondary ions (fragment ions) via CID in an MS/MS configuration utilizing, in particular, for the ionization and volatilization, the method of fast atomic bombardment (FAB mass spectrometry) or plasma desorption (PD mass spectrometry). As an example, the application of FAB to the analysis of protected dimeric blocks for chemical synthesis of oligodeoxynucleotides has been described (Köster et al., Biochemical Environmental Mass Spectrometry 14, 111-116 (1987)).

1. 2.8.2 Two more ionization/desorption techniques (ES and MALDI):

Two more recent ionization/desorption techniques are electrospray/ion spray (ES) and matrix-assisted laser desorption/ionization (MALDI). ES mass spectrometry has been introduced by Fenn et al. J. Phys. Chem. 18, 4451-59 (1984); PCT Application No. WO 90/14148) and current applications are summarized in recent review articles (R.D. Smith et al., Anal. Chem. 62, 882-89 (1990) and B. Ardrey, Electrospray Mass Spectrometry, Spectroscopy Europe 4, 10-18 (1992)). The molecular weights of the tetradecanucleotide d(CATGCCATGGCATG) (Covey et al. "The Determination of Protein, Oligonucleotide and Peptide Molecular Weights by Ion Spray Mass Spectrometry," Rapid Communications in Mass Spectrometry 2, 249-256 (1988)), of the 21-mer d(AAATTGTGCACATCCTGCAGC) and without giving details of that of a tRNA with 76 nucleotides Methods in Enzymology 1. 23, "Mass Spectrometry" (McCloskey, editor), p. 425, 1990, Academic Press, New York) have been published. As a mass analyzer, a quadrupole is most frequently used. The determination of molecular weights in femtomole amounts of sample is very accurate due to the presence of multiple ion peaks which all could be used for the mass calculation.

MALDI mass spectrometry, in contrast, can be particularly attractive when a time-of-flight (TOF) configuration is used as a mass analyzer. The MALDI-TOF mass spectrometry has been introduced by Hillenkamp et al. ("Matrix Assisted UV-Laser

Desorption/Ionization: A New Approach to Mass Spectrometry of Large Biomolecules, Biological Mass Spectrometry (Burlingame and McCloskey, editors), Elsevier Science Publishers, Amsterdam, pp. 49-60, 1990.) Since, in most cases, no multiple molecular ion peaks are produced with this technique, the mass spectra, in principle, look simpler compared to ES mass spectrometry. Although DNA molecules up to a molecular weight of 410,000 daltons could be desorbed and volatilized (Williams et al., "Volatilization of High Molecular Weight DNA by Pulsed Laser Ablation of Frozen Aqueous Solutions," Science, 246, 1585-87 (1989)), this technique has so far only been used to determine the molecular weights of relatively small oligonucleotides of known sequence, e.g., oligothymidylic acids up to 18 nucleotides (Huth-Fehre et al., "Matrix- Assisted Laser Desorption Mass Spectrometry of Oligodeoxythymidylic Acids," Rapid Communications in Mass Spectrometry, 6, 209-13 (1992)) and a double-stranded DNA of 28 base pairs (Williams et al., "Time-of-Flight Mass Spectrometry of Nucleic Acids by Laser Ablation and Ionization from a Frozen Aqueous Matrix," Rapid Communications in Mass Spectrometry, 4, 348-351 (1990)). In one publication (Huth- Fehre et al., 1992 , supra), it was shown that a mixture of all the oligothymidylic acids from n=12 to n=18 nucleotides could be resolved.

1. 2.8.3 Producing fragments, separating by electrophoresis and using matrix method to sequence

In U.S. Patent No. 5,064,754, RNA transcripts extended by DNA both of which are complementary to the DNA to be sequenced are prepared by incorporating NTP's, dNTP's and, as terminating nucleotides, ddNTP's which are substituted at the 5'- position of the sugar moiety with one or a combination of the isotopes ^{12}C , ^{13}C , ^{14}C , ^1H , ^2H , ^3H , ^{16}O , ^{17}O and ^{18}O . The polynucleotides obtained are degraded to 3'- nucleotides, cleaved at the N-glycosidic linkage and the isotopically labeled 5'- functionality removed by periodate oxidation and the resulting formaldehyde species determined by mass spectrometry. A specific combination of isotopes serves to discriminate base-specifically between internal nucleotides originating from the incorporation of NTPs and dNTP's and terminal nucleotides caused by linking ddNTP's to the end of the polynucleotide chain. A series of RNA/DNA fragments is produced, and in one embodiment, separated by electrophoresis, and, with the aid of the so-called matrix method of analysis, the sequence is deduced.

1. 2.8.4 Mass spectrometry using atoms which normally do not occur in DNA

In Japanese Patent No. 59-131909, an instrument is described which detects nucleic acid fragments separated either by electrophoresis, liquid chromatography or high speed gel filtration. Mass spectrometric detection is achieved by incorporating into the nucleic acids atoms which normally do not occur in DNA such as S, Br, I or Ag, Au, Pt, Os, Hg. The method, however, is not applied to sequencing of DNA using the Sanger method. In particular, it does not propose a base-specific correlation of such elements to an individual ddNTP.

1. 2.8.5 Sequencing with the Sanger method by using four stable isotopes to label the ddNTP's

PCT Application No. WO 89/12694 (Brennan et al., Proc. SPIE-Int. Soc. Opt. Eng. 1206, (New Technol. Cytom. Mot. Biol.), pp. 60-77 (1990); and Brennan, U.S. Patent No. 5,003,059) employs the Sanger methodology for DNA sequencing by using a combination of either the four stable isotopes ^{32}S , ^{33}S , ^{34}S , ^{36}S or ^{35}Cl , ^{37}Cl , ^{79}Br , ^{81}Br to specifically label the chain-terminating ddNTP's. The sulfur isotopes can be located either in the base or at the alpha-position of the triphosphate moiety whereas the halogen isotopes are located either at the base or at the 3'-position of the sugar ring.

The sequencing reaction mixtures are separated by an electrophoretic technique such as CZE, transferred to a combustion unit in which the sulfur isotopes of the incorporated ddNTP's are transformed at about 900°C in an oxygen atmosphere. The SO_2 generate with masses of 64, 65, 66 or 68 is determined on-line by mass spectrometry using, e.g., mass analyzer, a quadrupole with a single ion-multiplier to detect the ion current.

1. 2.8.6 Using resonance ionization spectroscopy in conjunction with a magnetic sector mass analyzer

A similar approach is proposed in U.S. Patent No. 5,002,868 (Jacobson et al., Proc. SPIE-Int. Soc. Opt. Eng. 1435, 9pt. Methods Ultrasensitive Detect. Anal. Tech. 26-35 (1991)) using Sanger sequencing with four ddNTP's specifically substituted at the alpha-position of the triphosphate moiety with one of the four stable sulfur isotopes as described above and subsequent separation of the four sets of nested sequences by tube gel electrophoresis. The only difference is the use of resonance ionization spectroscopy (RIS) in conjunction with a magnetic sector mass analyzer as

disclosed in U.S. Patent No. 4,442,354 to detect the sulfur isotopes corresponding to the specific nucleotide terminators, and by this, allowing the assignment of the DNA sequence.

1. 2.8.7 Using tube gel electrophoresis, a nebulizer and a mass analyzer to sequence

EPO Patent Applications No. 0360676 A1 and 0360677 A1 also describe Sanger sequencing using stable isotope substitutions in the ddNTP's such as D, ^{13}C , ^{15}N , ^{17}O , ^{18}O , ^{32}S , ^{33}S , ^{34}S , ^{36}S , ^{19}F , ^{35}Cl , ^{37}Cl , ^{79}Br , ^{81}Br and ^{127}I or function groups such as CF_3 or $\text{Si}(\text{CH}_3)_3$ at the base, the sugar or the alpha position of the triphosphate moiety according to chemical functionality. The Sanger sequencing reaction mixtures are separated by tube gel electrophoresis. The effluent is converted into an aerosol by the electrospray/thermospray nebulizer method and then atomized and ionized by a hot plasma (7000 to 8000°K) and analyzed by a simple mass analyzer. An instrument is proposed which enables one to automate the analysis of the Sanger sequencing reaction mixture consisting of tube electrophoresis, a nebulizer and a mass analyzer.

The application of mass spectrometry to perform DNA sequencing by the hybridization/fragment method (see above) has been recently suggested (Bains, "DNA Sequencing by Mass Spectrometry: Outline of a Potential Future Application, *Chimicaoggi* 2, 13-16 (1991)).

1. 2.9 Probes

1. 2.9.1 Using large arrays of nucleic acid probes on a substrate

Alternative techniques have been proposed for sequencing a nucleic acid. PCT patent Publication No. 92/10588, incorporated herein by reference for all purposes, describes one improved technique in which the sequence of a labeled, target nucleic acid is determined by hybridization to an array of nucleic acid probes on a substrate. Each probe is located at a positionally distinguishable location on the substrate. When the labeled target is exposed to the substrate, it binds at locations that contain complementary nucleotide sequences. Through knowledge of the sequence of the probes at the binding locations, one can determine the nucleotide sequence of the target nucleic acid. The technique is particularly efficient when very large arrays of nucleic acid probes are utilized.

Such arrays can be formed according to the techniques described in U.S. Patent No. 5,143,854 issued to Pirrung et al. See also U.S. application Serial No. 07/805,727, both incorporated herein by reference for all purposes.

1. 2.9.2 Employing sequencing by hybridization when the probes are shorter than the target

When the nucleic acid probes are of a length shorter than the target, one can employ a reconstruction technique to determine the sequence of the larger target based on affinity data from the shorter probes. See U.S. Patent No. 5,202,231 to Drmanac-et al., and PCT patent Publication No. 89/10977 to Southern. One technique for overcoming this difficulty has been termed sequencing by hybridization or SBH. For example, assume that a 12-mer target DNA 5'-AGCCTAGCTGAA is mixed with an array of all octanucleotide probes. If the target binds only to those probes having an exactly complementary nucleotide sequence, only five of the 65,536 octamer probes (3'-TCGGATCG, CGGATCGA, GGATCGAC, GATCGACT, and ATCGACTT) will hybridize to the target. Alignment of the overlapping sequences from the hybridizing probes reconstructs the complement of the original 12-mer target:

TCGGATCG
CGGATCGA
GGATCGAC

GATCGACT
ATCGACTT
TCGGATCGACTT

While meeting with much optimism, prior techniques have also met with certain limitations. For example, practitioners have encountered substantial difficulty in analyzing probe arrays hybridized to a target nucleic acid due to the hybridization of partially mismatched sequences, among other difficulties. The present invention provides significant advances in sequencing with such arrays.

1. 2.10. DNA Amplification

DNA can be amplified by a variety of procedures including cloning (Sambrook et al., *Molecular Cloning : A Laboratory Manual*, Cold Spring Harbor Laboratory Press, 1989), polymerase chain reaction (PCR) (C.R. Newton and A. Graham, PCF, BIOS Publishers, 1994), ligase chain reaction (LCR) (F. Barany Proc. Natl. Acad Sci USA 88, 189-93 (1991), strand displacement amplification (SDA) (G. Terrance Walker et al., *Nucleic Acids Res.* 22, 2670-77 (1994)) and variations such as RT-PCR, allele-specific amplification (ASA) etc.

The polymerase chain reaction (Mullis, K. et al., *Methods Enzymol.*, 155:335-350 1987) permits the selective in vitro amplification of a particular DNA region by mimicking the phenomena of in vivo DNA replication. Required reaction components are single stranded DNA, primers (oligonucleotide sequences complementary to the 5' and 3' ends of a defined sequence of the DNA template), deoxynucleotidetriphosphates and a DNA polymerase enzyme. Typically, the single stranded DNA is generated by heat denaturation of provided double strand DNA. The reaction buffers contain magnesium ions and co-solvents for optimum enzyme stability and activity.

The amplification results from a repetition of such cycles in the following manner: The two different primers, which bind selectively each to one of the complementary strands, are extended in the first cycle of amplification. Each newly synthesized DNA then contains a binding site for the other primer. Therefore each new DNA strand becomes a template for any further cycle of amplification enlarging the template pool from cycle to cycle. Repeated cycles theoretically lead to exponential synthesis of a DNA-fragment with a length defined by the 5' termini of the primer.

The PCR amplification procedure has been used to sequence the DNA being amplified (e.g. "Introduction to the AmpliTaq Cycle Sequencing Kit Protocol", a booklet from Perkin Elmer Cetus Corporation). The DNA could be first amplified and then it could be sequenced using the two conventional DNA sequencing techniques. Modified methods for sequencing PCR-amplified DNA have also been developed (e.g. Bevan et al., "Sequencing of PCR-Amplified DNA" *PCR Meth. App.* 4:222 (1992)).

1. 2.11 Additional Sequencing Methods

1. 2.11.1 Sanger sequencing using the degradation of phosphorothioate-containing DNA fragments

A recent modification of the Sanger sequencing strategy involves the degradation of phosphorothioate-containing DNA fragments obtained by using alpha-thio dNTP instead of the normally used ddNTPs during the primer extension reaction mediated by DNA polymerase (Labeit et al., MA 5, 173-177 (1986); Amersham, PCT- Application GB86/00349; Eckstein et al., Nucleic Acids Res. 1~, 9947 (1988)). Here, the four sets of base-specific sequencing ladders are obtained by limited digestion with exonuclease III or snake venom phosphodiesterase, subsequent separation on PAGE and visualization by radioisotopic labeling of either the primer or one of the dNTPs. In a further modification, the base-specific cleavage is achieved by alkylating the sulphur atom in the modified phosphodiester bond followed by a heat treatment (Max- Planck- Gesellschaft, DE 3930312 A1). Both methods can be combined with the amplification of the DNA via the Polymerase Chain Reaction (PCR).

1. 2.11.2 Sanger sequencing using modified polymerization reaction (at high temperature)

Initial PCR experiments used thermolabile DNA polymerase. However, thermolabile DNA polymerase must be continually added to the reaction mixture after each denaturation cycle. Major advances in PCR practice were the development of a polymerase, which is stable at the near-boiling temperature (Saiki, R. et al., Science 239:487-491 1998) and the development of automated thermal cyclers.

The discovery of thermostable polymerases also allowed modification of the Sanger sequencing reaction with significant advantages. The polymerization reaction could be carried out at high temperature with the use of thermostable DNA polymerase in a cyclic manner (cycle sequencing). The conditions of the cycles are similar to those of the PCR technique and comprise denaturation, annealing, and extension steps. Depending on the length of the primers only one annealing step at the beginning of the reaction may be sufficient. Carrying out a sequencing reaction at high temperature in a cyclic manner provides the advantage that each DNA strand can serve as template in every new cycle of extension which reduces the amount of DNA

necessary for sequencing, thereby providing access to minimal volumes of DNA, as well as resulting in improved specificity of primer hybridization at higher temperature and the reduction of secondary structures of the template strand.

1. 2.11.3 Semi-exponential cycle sequencing using a second reverse primer in the sequencing reaction

However, amplification of the terminated fragments is linear in conventional cycle sequencing approaches. A recently developed method, called semi-exponential cycle sequencing shortens the time required and increases the extent of amplification obtained from conventional cycle sequencing by using a second reverse primer in the sequencing reaction. However, the reverse primer only generates additional template strands if it avoids being terminated prior to reaching the sequencing primer binding site. Needless to say, terminated fragments generated by the reverse primer can not serve as a sufficient template. Therefore, in practice, amplification by the semi-exponential approach is not entirely exponential. (Sarkat, G. and Bolander Mark E., Semi Exponential Cycle Sequencing Nucleic Acids Research, 1995, Vol. 23, No. 7, p. 1269-1270).

1. 2.11.4 Need to facilitate highthroughput sequencing

In addition to the foregoing limitations inherent in current sequencing techniques, the generation of DNA substrate molecules for each 300 to 500 nucleotides to be sequenced is presently required. Assuming no overlapping sequence between substrate molecules, the sequencing of both strands of an entire mammalian genome would, therefore, require the generation of at least 20 million DNA substrate molecules.

As pointed out above, current nucleic acid sequencing methods require relatively large amounts (typically about 1 g) of highly purified DNA template. Often, however, only a small amount of template DNA is available. Although amplifications may be performed, amplification procedures are typically time consuming, can be limited in the amount of amplified template produced and the amplified DNA must be purified prior to sequencing. A streamlined process for amplifying and sequencing DNA is needed, particularly to facilitate highthroughput nucleic acid sequencing.

1. 2.12 Strategies for obtaining the initial sequence

Methods currently used to sequence large segments of DNA do not lend themselves to large-scale determination of genomic sequences. In general, the initial determination of a genomic clone sequence results in ambiguities and discrepancies that are resolved by assembling and editing the raw sequencing data into a consensus sequence. There are also, generally, holes in the sequence that need to be filled in in order to create a finished sequence. There are two general strategies for obtaining the initial sequence: shotgun sequencing and transposon-mediated directed sequencing.

1. 2.12.1 Shotgun sequencing

In the currently existing methods for sequencing very long DNA of millions of nucleotides, the DNA is fragmented into smaller, overlapping fragments, and sub-cloned to produce numerous clones containing overlapping DNA sequences. These clones are sequenced randomly and the sequences assembled by "overlap sequence-matching" to produce the contiguous sequence. In this shot-gun sequencing method, approx. ten times more sequencing than the length of the DNA being sequenced is required to assemble the contiguous sequence. Shotgun sequencing is reasonably appropriate for generating the initial sequences of the genomic clone. In this method, the clone is digested with a multiplicity of restriction enzymes and the individual fragments are sequenced. When sufficient sequence is obtained to putatively cover the length of the genomic clone (1 x total sequence length) statistically 65% of the genomic clone sequence will have successfully been determined. The shotgun strategy relies on assembly algorithms to piece together a final sequence by determining relationships between a selected set of random templates. Although this assembly process is semiautomated, it remains labor-intensive, especially in complex regions that contain highly related tandem repeats. In addition, since the selection of subclones is not random, gaps of unknown distance are included between islands of known sequence. Linking up the islands requires either sequencing additional subclones or ordering custom oligonucleotides to generate sequence into the gaps. The weaknesses of shotgun sequencing performed on substantial lengths of nucleotide sequence are thus 1) the difficulties involved in sequence assembly and 2) the need for hole-filling.

A non-ordered approach to sequencing, e.g., shotgun sequencing, would require the generation of 100 to 200 million DNA templates. Although there has been

effort directed to automating the steps presently involved in DNA substrate generation, e.g., restriction mapping, preparation of subfragments for subcloning, identification of subclones, growing bacterial cultures, and purifying nucleic acids, it is unlikely that human intervention can be substantially eliminated from the process. Current approaches, therefore, are less than optimal for the large scale sequencing of DNA, particularly sequencing the human genome.

Although the problems enumerated above are not intended to be exhaustive, the limitations inherent in methods presently available for sequencing DNA are readily apparent. Accordingly, there exists a need for an improved method of sequencing DNA that circumvents the need for primer binding sites as well as the need to determine restriction maps. Additionally, there exists a need for an improved method which extends the amount of sequence information obtainable from a DNA substrate, thus substantially reducing the number of DNA substrate molecules required to sequence a given region of DNA. The present invention meets these needs.

1.2.12.2 Transposon-mediated directed sequencing

On the other hand, the transposon-mediated sequencing method described by Strathmann, M. et al. Proc Natl Acad Sci USA (1991) 88:1247- 1250, provides an orderly approach to generating subclones for sequencing. The method uses a γ - δ bacterial transposable element bracketed by sequencing primers. The primer-flanked transposon permits the introduction of evenly spaced priming sites across a fragment with an unknown DNA sequence. The number of template sequences required to obtain the complete sequence information can be calculated from the length of the fragment. In the "directed" sequencing method, the linear order of the DNA clones has to be first determined by "physical mapping" of the clones. As the transposon insertions are random, the positions of the insertions are mapped, for example, using the polymerase chain reaction (PCR) using primers that amplify the intervening sequence between the transposon insertion site and the vector sequences at each end of the inserted fragment to be sequenced. The lengths of the amplified products thus define a map position for the transposon. Sequencing can be conducted based on the sequencing primers flanking the transposon, and since the position of the transposon has been mapped prior to sequencing, a fully automated assembly process

is possible. There are no gaps since an ordered set of sequencing templates which cover the DNA fragment is produced.

1.2.12.3 Drawbacks of these two strategies, "primer-walking" method

However, transposon sequencing can only be used on fragments containing 2-5 kb; preferably 3-4 kb. Thus, to use the transposon method on larger fragments, smaller subclones of the original fragment must be generated and organized into an ordered overlapping set. The shotgun strategy is not completely appropriate for this purpose. Neither is an alternative strategy termed dog-tagging. Dog-tagging is a "walking" process, a contiguous DNA sequencing method called the "primer-walking" method using the Sanger's DNA polymerase enzymatic sequencing procedure, that scans through a 30-hit subclone library for sequences that are near the end of the last walking step. It is labor-intensive and does not always succeed. In this method, the DNA copying has to occur always from the template DNA during DNA sequencing. In contrast, in the PCR procedure, the target DNA amplified in the first rounds from the original input template DNA will function as the template DNA in subsequent cycles of amplification. After a certain cycles of amplification, the DNA sequencing reaction will be started by adding the sequencing "cocktail". Thus in the PCR reaction, only one copy of template DNA is theoretically sufficient to amplify into millions of copies, and therefore a very little genomic (or template) DNA is sufficient for sequencing. The advantage of DNA amplification that exists in PCR is lacking in the conventional Sanger procedure. Thus, this primer-walking method will require a larger amount of template DNA compared to the PCR sequencing method. Also, because the long DNA has a tendency to re-anneal back to duplex DNA, the sequencing gel pattern may not be as clean as in a PCR procedure, when a very long DNA is being sequenced. This may limit the length of the DNA, that could be contiguously sequenced without breaking the DNA, using the primer-walking procedure. The PCR method also enables the reduction of non-specific binding of the primers to the template DNA because the enzymes used in these protocols function at high-temperatures, and thus allow "stringent" reaction conditions to be used to improve sequencing.

The present method of contiguous DNA sequencing using the basic PCR technique has thus many advantages over the primer walking method. Also, so far no method exists for contiguously sequencing a very long DNA using PCR technique.

The present invention thus offers a unique and very advantageous procedure for contiguous DNA sequencing.

1. 2.12.4 Amplification and sequencing a long genomic DNA without subcloning into smaller fragments

In one embodiment, the present invention provides a method for contiguous sequencing of very long DNA using a modification of the standard PCR technique without the need for breaking down and subcloning the long DNA.

The PCR technique enables the amplification of DNA which lies between two regions of known sequence (K. B. Mullis et al., U.S. Pat. Nos. 4,683,202; 7/1987; 435/91; and 4,683,195, 7/1987; 435/6). Oligonucleotides complementary to these known sequences at both ends serve as "primers" in the PCR procedure. Double stranded target DNA is first melted to separate the DNA strands, and then oligonucleotide (oligo) primers complementary to the ends of the segment which is desired to be amplified are annealed to the template DNA. The oligos serve as primers for the synthesis of new complementary DNA strands, using a DNA polymerase enzyme and a process known as primer extension. The orientation of the primers with respect to one another is such that the 5' to 3' extension product from each primer contains, when extended far enough, the sequence which is complementary to the other oligo. Thus, each newly synthesized DNA strand becomes a template for synthesis of another DNA strand beginning with the other oligo as primer. Repeated cycles of melting, annealing of oligo primers, and primer extension lead to a (near) doubling, with each cycle, of DNA strands containing the sequence of the template beginning with the sequence of one oligo and ending with the sequence of the other oligo.

The key requirement for this exponential increase of template DNA is the two oligo primers complementary to the ends of the sequence desired to be amplified, and oriented such that their 3' extension products proceed toward each other. If the sequence at both ends of the segment to be amplified is not known, complementary oligos cannot be made and standard PCR cannot be performed. The object of the present invention is to overcome the need for sequence information at both ends of the segment to be amplified, i.e. to provide a method which allows PCR to be performed when sequence is known for only a single region, and to provide a method for the

contiguous sequencing of a very long DNA without the need for subcloning of the DNA.

Amplifying and sequencing using the PCR procedure requires that the sequences at the ends of the DNA (the two primer sequences) be known in advance. Thus, this procedure is limited in utility, and cannot be extended to contiguously sequence a long DNA strand. If the knowledge of only one primer is sufficient without anything known about the other primer, it would be greatly advantageous for sequencing very long DNA molecules using the PCR procedure. It would then be possible to use such a method for contiguously sequencing a long genomic DNA without the need for subcloning it into smaller fragments, and knowing only the very first, beginning primer in the whole long DNA.

1. 2.12.5 Large-scale sequencing through the generation of a subclone path

In another embodiment, the present invention provides a large-scale sequencing method which combines efficient method to generate a subclone path through the large original fragment, such as a genomic clone, wherein the subclones are accessible to transposon sequencing, in combination with sequencing these subclones using the transposon method.

1. 2.13 Constructing ordered clone maps of DNA sequences

A primary goal of the human genome project is to determine the entire DNA sequence for the genomes of human, model, and other useful organisms. A related goal is to construct ordered clone maps of DNA sequences at 100 kilobase (kb) resolution for these organisms (D. R. Cox, E. D. Green, E. S. Lander, D. Cohen, and R. M. Myers, "Assessing mapping progress in the Human Genome Project," *Science*, vol. 265, no. 5181, pp. 2031-2, 1994), incorporated by reference. Integrated maps that localize clones together with polymorphic genetic markers (J. Weber and P. May, "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction," *Am. J. Hum. Genet.*, vol. 44, pp. 388-396, 1989), incorporated by reference, are particularly useful for positionally cloning human disease genes (F. Collins, "Positional cloning: lets not call it reverse anymore," *Nature Genet.*, vol. 1, no. 1, pp. 3-6, 1992), incorporated by reference. The greatest need, however, is for sequence-ready maps. Also useful are maps of expressed sequences. Mapping techniques include restriction enzyme analysis of genetic material, and the hybridization and detection of specific oligonucleotides which test for the presence or absence of particular alleles or loci, and may further be used to gain spatial information about the occurrence of their targets when appropriate analytic techniques are subsequently applied. Note that such characterizations presently are methodologically and operationally distinct from other processes comprehended within the biotechnological and related arts. Human DNA sequences now exist as genomic libraries in a variety of small- and large-insert capacity cloning vectors, with yeast artificial chromosomes (YACs) (D. T. Burke, G. F. Carle, and M. V. Olson, "Cloning of large exogenous DNA into yeast by means of artificial chromosomes," *Science*, vol. 236, pp. 806-812, 1987), incorporated by reference, used extensively in mapping large regions. Efficient strategies for performing the requisite experimentation are critical for sequencing and mapping chromosomes or entire genomes.

1. 2.13.1 Sequence-tagged site

The starting point for an effective sequencing method is a complete ordered clone map of a genome. Current strategies for ordering clones build contiguous sequences (contigs) using short-range comparison data. Sequence-tagged site (STS) (M. Olson, L. Hood, C. Cantor, and D. Botstein, "A common language for physical

mapping of the human genome," *Science*, vol. 245, pp. 1434-35, 1989), incorporated by reference, comparisons with clones are used in STS-content mapping (SCM) (E. D. Green and P. Green, "Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences," *PCR Methods and Applications*, vol. 1, pp. 77-90, 1991), incorporated by reference. For chromosomal or genome-wide SCM, very large YACs (megaYACs) are required for the currently available STS densities (R. Arratia, E. S. Lander, S. Tavaré, and M. S. Waterman, "Genomic mapping by anchoring random clones: a mathematical analysis," *Genomics*, vol. 11, pp. 806-827, 1991; W. J. Ewens, C. J. Bell, P. J. Donnelly, P. Dunn, E. Matallana, and J. R. Ecker, "Genome mapping with anchored clones: theoretical aspects," *Genomics*, vol. 11, pp. 799-805, 1991), incorporated by reference; these large YACs are often chimeric or contain gaps. Restriction fragment fingerprint mapping has been done with hybridization (C. Bellanne-Chantelot, B. Lacroix, P. Ougen, A. Billault, S. Beaufils, S. Bertrand, S. Georges, F. Glibert, I. Gros, G. Lucotte, L. Susini, J.-J. Codani, P. Gesnouin, S. Pook, G. Vaysseix, J. Lu-Kuo, T. Ried, D. Ward, I. Chumakov, D. Le Paslier, E. Barillot, and D. Cohen, "Mapping the whole genome by fingerprinting yeast artificial chromosomes," *Cell*, vol. 70, pp. 1059-1068, 1992; R. L. Stallings, D. C. Torney, C. E. Hildebrand, J. L. Longmire, L. L. Deaven, J. H. Jett, N. A. Doggert, and R. K. Moyzis, "Physical mapping of human chromosomes by repetitive sequence hybridization," *Proc. Natl. Acad. Sci. USA*, vol. 87, pp. 6218-6222, 1990), incorporated by reference, or without hybridization (A. Coulson, J. Sulston, S. Brenner, and J. Karn, "Toward a physical map of the genome of the nematode *Caenorhabditis elegans*," *Proc. Natl. Acad. Sci. USA*, vol. 83, pp. 7821-7825, 1986), incorporated by reference. With hybridization fingerprinting, path analysis of YAC fingerprints is not always reliable when constructing contigs. Hybridizing an internal clone sequence (e.g., end-clone sequence, Alu-PCR probes) against a library to determine neighboring sequences builds unpositioned YAC contigs (M. T. Ross and V. P. J. Stanton, "Screening large-insert libraries by hybridization," in *Current Protocols in Human Genetics*, vol. 1, N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed. New York: John Wiley and Sons, 1995, pp. 5.6.1-5.6.34), incorporated by reference, although walking techniques are generally reserved for closing gaps.

1. 2.13.2 Gridding library onto nylon filters, and hybridizing with probes to reduce cost, increase throughput

The number of experiments needed for these short-range clone mapping approaches increases with the number of clones in the library. While considerable efficiency is gained by using multiplexed experiments with pooled reagents (G. A. Evans and K. A. Lewis, "Physical mapping of complex genomes by cosmid multiplex analysis," *Proc. Natl. Acad. Sci. USA*, vol. 86, no. 13, pp. 5030-4, 1989; E. D. Green and M. V. Olson, "Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction," *Proc. Natl. Acad. Sci. USA*, vol. 87, no. 3, pp. 1213-7, 1990), incorporated by reference, the experimental requirements are at least proportional to the number of clones. A useful goal is to significantly reduce cost and increase throughput by achieving a number of required experiments largely independent of library size. One step toward this independence has been achieved by gridding an entire library onto nylon filters, and then hybridizing these filters with a set of probes (H. Lehrach, A. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A. P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka, "Hybridization fingerprinting in genome mapping and sequencing," in *Genetic and Physical Mapping I: Genome Analysis*, K. E. Davies and S. M. Tilghman, ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory, 1990, pp. 39-81; A. P. Monaco, V. M. S. Lam, G. Zehetner, G. G. Lennon, C. Douglas, D. Nizetic, P. N. Goodfellow, and H. Lehrach, "Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products," *Nucleic Acids Res.*, vol. 19, no. 12, pp. 3315-3318, 1991), incorporated by reference. For example, contigs of small genomic regions have been constructed by oligonucleotide fingerprinting of gridded cosmid filters (A. G. Craig, D. Nizetic, J. D. Hoheisel, G. Zehetner, and H. Lehrach, "Ordering of cosmid clones covering the herpes simplex virus type I," *Nucleic Acids Res.*, vol. 18, no. 9, pp. 2653-60, 1990; A. J. Cuticchia, J. Arnold, and W. E. Timberlake, "ODS: ordering DNA sequences, a physical mapping algorithm based on simulated annealing," *CABIOS*, vol. 9, no. 2, pp. 215-219, 1992), incorporated by reference.

1. 2.13.3 Radiation hybrid mapping

To efficiently span larger genomic regions, radiation hybrid (RH) mapping (D. R. Cox, M. Burmeister, E. R. Price, S. Kim, and R. M. Myers, "Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes," *Science*, vol. 250, pp. 245-250, 1990), incorporated by reference, has been used to localize small DNA sequences (though not clones) into high-resolution bins. Relatively few PCR experiments with one 96-well plate library of RHs generally suffice for mapping STSs or genes to unique bins having 250 kb to 1 Mb average resolution. The very large multiple fragments in each RH clone efficiently cover much of a chromosome (or genome). Assaying a sequence for intersection against a set of RHs provides long-range relational information for localization much akin to somatic cell hybrid (SCH) mapping (M. C. Weiss and H. Green, "Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes," *Proc. Natl. Acad. Sci. USA*, vol. 58, pp. 1104-1111, 1976), incorporated by reference. However, RH mapping offers much greater resolution than SCH or fluorescent in situ hybridization (FISH) mapping.

1. 2.13.4 Combining RH mapping with filter hybridization techniques

For highly optimized experimentation, it would be desirable to combine high-resolution long-range RH mapping with low-cost high-throughput filter hybridization techniques to map clones. One can serially probe a gridded clone library with a set of RHs (H. Lehrach, A. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A. P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka, "Hybridization fingerprinting in genome mapping and sequencing," in *Genetic and Physical Mapping I: Genome Analysis*, K. E. Davies and S. M. Tilghman, ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1990, pp. 39-81), in principle requiring a number of experiments that is independent of the clone library size and logarithmically related to the desired map resolution. However, complex hybridization probes such as RHs (or their Alu-PCR products) generate data containing considerable noise. This inherent uncertainty, together with the large clone insert size (which complicates conventional RH analysis), has thus far precluded high-resolution mapping of clones using RHs (J. Kumlien, T. Labella, G. Zehetner, R. Vatcheva, D. Nizetic, and H. Lehrach, "Efficient identification and regional positioning of YAC and cosmid clones to human

chromosome 21 by radiation fusion hybrids," *Mammalian Genome*, vol. 5, no. 6, pp. 365-71, 1994), incorporated by reference.

1. 2.13.5 Inner product mapping

Inner product mapping (IPM) is a hybridization-based method for achieving high-throughput, high-resolution RH mapping of clones (M. W. Perlin and A. Chakravarti, "Efficient construction of high-resolution physical maps from yeast artificial chromosomes using radiation hybrids: inner product mapping," *Genomics*, vol. 18, pp. 283-289, 1993), incorporated by reference, that overcomes this barrier. Experimental data have established that IPM is a highly rapid, inexpensive, accurate, and precise large-scale long-range mapping method, particularly when preexisting RH maps are available, and that IPM can replace or complement more conventional short-range mapping methods.

1. 2.13.6 Obtaining improved mapping results

Improved mapping results can be obtained incrementally by gradually enlarging the data tables, a process which provides useful feedback to both experimentation and analysis. With additional RHs, the signal-to-noise characteristics of the clone profiles improve. This incremental process, and the relatively few RHs required for accurate mapping, follows the logarithmic number of the probes needed for IPM. For best mapping results, as many STS-typed RHs as feasible are used: with currently available high-throughput, robotically-assisted hybridization methods, the localization benefits of performing many filter hybridizations outweigh the relatively low experimentation costs. The incremental construction also highlights IPM's indirect inference of map location: STS-content mapping directly compares clones with STSs, and can not map small-insert clones against STSs which are insufficiently dense.

1. 2.13.7 Building accurate maps and partitioning data noise

IPM builds accurate maps from low-confidence data. IPM's partitioning of the experiments into two data tables of (A) clones vs. RHs and (B) RHs vs. STSs also partitions the data noise. Table B is formed from relatively noiseless PCR-based comparisons of STSs against RH DNA, and can thus accurately order and position the STS bins using combinatorial mapping procedures (M. Boehnke, "Radiation hybrid mapping by minimization of the number of obligate chromosome breaks," *Genetic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes*.

Cytogenet Cell Genet, vol. 59, pp. 96-98, 1992; M. Boehnke, K. Lange, and D. R. Cox, "Statistical methods for multipoint radiation hybrid mapping," Am. J. Hum. Genet., vol. 49, pp. 1174-1188, 1991), incorporated by reference. Table A is formed from inherently unreliable and inconsistently replicated hybridizations of complex RH probes against gridded filters. Inner product mapping uses the table B data matrix to ameliorate these data errors and robustly translate a clones's noisy RH signature vector (a row of table A) into a chromosomal profile, whose peak bins the clone.

1. 2.13.8 Mapping YAC's using IPM

IPM is a proven approach for mapping YACs (C. W. Richard III, D. J. Duggan, K. Davis, J. E. Farr, M. J. Higgins, S. Qin, L. Zhang, T. B. Shows, M. R. James, and M. W. Perlin, "Rapid construction of physical maps using inner product mapping: YAC coverage of chromosome 11," in Fourth International Conference on Human Chromosome 11, Sep. 22-24, Oxford, England, 1994), incorporated by reference, and is a candidate method for mapping PACs (P. A. Ioannou, C. T. Amemiya, J. Barnes, P. M. Kroisel, H. Shizuya, C. Chen, M. A. Batzer, and P. J. de Jong, "A new bacteriophage P1-derived vector for the propagation of large human DNA fragments," Nature Genet., vol. 6, no. 1, pp. 84-89, 1994), incorporated by reference, cosmids, expressed sequences (M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merrill, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter, "Complementary DNA sequencing: Expressed sequence tags and human genome project," Science, vol. 252, pp. 1651-1656, 1991), incorporated by reference, and other physical reagents (J. D. McPherson, C. Wagner- McPherson, M. Perlin, and J. J. Wasmuth, "A physical map of human chromosome 5 (Abstract)," Amer. J. Hum. Genet., vol. 55, no. 3 Supplement, pp. A265, 1994), incorporated by reference. Hybridization efficiency for table A can be improved by using long and IRE-bubble PCR (D. J. Munroe, M. Haas, E. Bric, T. Whirton, H. Aburatani, K. Hunter, D. Ward, and D. E. Housman, "IRE-bubble PCR: a rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources," Genomics, vol. 19, no. 3, pp. 506-14, 1994), incorporated by reference, to reduce false negative errors, providing controls and redundant DNA spotting for internal calibration, and directly acquiring signals (e.g., via a phosphorimager, Molecular Dynamics, Sunnyvale, Calif.) to facilitate automated scoring. Current robotic technologies enable the high-throughput

construction of gridded filters (A. Copeland and G. Lennon, "Rapid arrayed filter production using the 'ORCA' robot," *Nature*, vol. 369, no. 6479, pp. 421-422, 1994), incorporated by reference; single use of these filters would reduce the time and error related to stripping and reprobing. Robots similarly provide high-throughput PCR comparisons for constructing table B. Alternatively, existing RH mapping data can be rapidly extended (at low cost) into inner product maps of libraries (U. Francke, E. Chang, K. Comeau, E.-M. Geigl, J. Giacalone, X. Li, J. Luna, A. Moon, S. Welch, and P. Wilgenbus, "A radiation hybrid map of human chromosome 18," *Cytogenet. Cell Genet.*, vol. 66, pp. 196-213, 1994), incorporated by reference.

1. 2.13.9 Whole genome RH libraries

Whole human genome RH (WG-RH) libraries of 0.5 and 1.0 Mb resolution have been constructed (D. R. Cox, K. O'Connor, S. Hebert, M. Harris, R. Lee, B. Stewart, G. DiSibio, M. Boehnke, K. Lange, R. Goold, and R. M. Myers, "Construction and analysis of a panel of 'whole genome' radiation hybrids (Abstract)," *Amer. J. Hum. Genet.*, vol. 55, no. 3 Supplement, pp. A23, 1994; M. A. Walter, D. J. Spillerr, P. Thomas, J. Weissenbach, and P. N. Goodfellow, "A method for constructing radiation hybrid maps of whole genomes," *Nature Genet.*, vol. 7, no. 1, pp. 22-28, 1994), incorporated by reference, and have been characterized for the STSs used in the genome-wide CEPH megaYAC STS-content map (T. Hudson, S. Foote, S. Gerety, J. Ma, S.-h. Xu, X. Hu, J. Bae, J. Silva, J. Valle, S. Maitra, A. Colbert, L. Horton, M. Anderson, M. P. Reeve, M. Daly, A. Kaufman, C. Rosenberg, L. Stein, N. Goodman, J. Orlin, D. C. Page, and E. S. Lander, "Towards an STS-content map of the human genome (Abstract)," *Amer. J. Hum. Genet.*, vol. 55, no. 3 Supplement, pp. A23, 1994), incorporated by reference. The availability of this WG-RH table B resource suggests that constructing table A by performing hybridizations between species specific (e.g., Alu-PCR) products of these RHs and gridded clones or expressed sequences, and then combining tables A and B to build a genome-wide inner product map, is a fast, accurate, and inexpensive approach to whole genome physical mapping. IPM has localized the components of chimeric YACs as distinct multiple peaks. IPM is therefore useful in verifying and extending current megaYAC mapping projects, and in multiplexed experimental designs that pool sequences from well-separated bins.

1. 2.13.10 Using short-range data to determine the orders and distances of clone subsets in proximate bins

IPM provides long-range mapping information for DNA sequences relative to RH bins through DNA hybridization. This binning information can be complemented with short-range mapping data, such as oligonucleotide fingerprint hybridizations (H. Lehrach, A. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A. P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka, "Hybridization fingerprinting in genome mapping and sequencing," in *Genetic and Physical Mapping I: Genome Analysis*, K. E. Davies and S. M. Tilghman, ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory, 1990, pp. 39-81), incorporated by reference, and (R. Drmanac, Z. Strezoska, I. Labat, S. Drmanac, and R. Crkvenjakov, "Reliable hybridization of oligonucleotides as short as six nucleotides," *DNA Cell Biol.*, vol. 9, no. 7, pp. 527-534, 1990), incorporated by reference. Combining the data from these two high-throughput hybridization studies enables a two-pass BIN-SORT (A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*. Reading, Mass.: Addison-Wesley, 1983), incorporated by reference, strategy to high-resolution mapping: first use IPM to bin the clones, and then use short-range data to determine the orders and distances of clone subsets in proximate bins. This strategy can rapidly construct minimum-length paths of sequence-ready clones that tile the genome. Crucially, such IPM-derived contigs overcome the short-range limitations of all other known mapping methods, and enable the coordinated sequencing of the human genome, which is a well-recognized goal (F. Collins and D. Galas, "A new five-year plan for the U.S. Human Genome Project," *Science*, vol. 262, pp. 43-46, 1993), incorporated by reference. Such combination approaches can be highly effective for other purposes, such as using short-range proximity data to sharpen long-range inner product map results. IPM's experimental efficiencies enable effective determination of genome-wide DNA sequences, and the construction of high-resolution integrated genome maps for human, model organism, and agricultural species.

In one embodiment, this invention pertains to determining the sequence of the genome of an organism or species through the use of a novel, unobvious, and highly effective clone mapping strategy. Such sequence information can be used for finding genes of known utility, determining structure/function properties of genes and their products, elucidating metabolic networks, understanding the growth and development

of humans and other organisms, and making comparisons of genetic information between species. From these studies, diagnostic tests and pharmacological agents can be developed of great utility for preventing and treating human and other disease.

Disclosures of this type, yielded in a search, are:

Patent Number: Inventor Issued* US 5,302,509 Cheeseman, Peter C. 12 April 1994 WO 93/2134 Rosenthal., A; et al. 28 October 1993 DE 41 41 178 Al Ansorge, Wilhelm 16 June 1993 Wo 93/01583 Gibbs, Richard A.; et al. 18 March 1993 Wo 91/06678 Tsien, Roger Y.; et al. 16 May 1991 WO 90/13666 Garland, Peter B.; et al. 15 November 1990 Included in some of these above disclosures are descriptions of nucleotide triphosphates comprising removable fluorescent 3' protecting groups.

1.3 ALTERNATIVE SEQUENCING METHODS

The present invention provides an improved method of determining the nucleotide base sequence of DNA. In one embodiment, the method of the invention involves the preparation of a DNA substrate comprising a set of molecules, each having a template strand and a primer strand, wherein the 3' ends of the primer strands of the molecules terminate at about the same nucleotide position on the template strands of the molecules within each set. Preferably, the template and primer strands of the molecules are of unequal lengths wherein the 3' ends of the primer strands of the molecules terminate at about the same nucleotide position on the template strands of the molecules within each set. DNA synthesis is induced to obtain labeled reaction products comprising newly synthesized DNA complementary to the template strands using the 3' ends of the primer strands to prime DNA synthesis, labeled nucleoside triphosphates, at least one modified nucleoside triphosphate, and preferably, a suitable chain terminator, wherein the modified nucleoside triphosphate is selected to substantially protect newly synthesized DNA from cleavage. Thereafter, the labeled reaction products are cleaved at one or more selected sites to obtain labeled DNA fragments wherein newly synthesized DNA is substantially protected from cleavage by the incorporation of the modified nucleotide. The labeled DNA fragments obtained in the preceding step are separated and their nucleotide base sequence is identified by suitable means. The advantages of the present invention over prior art methods will become apparent after consideration of the accompanying drawings and the following detailed description of the invention.

1.3.1 One-step process for generating from a DNA template

According to one process of the invention, a combined amplification and termination reaction is performed using at least two different polymerase enzymes, each having a different affinity for the chain terminating nucleotide, so that polymerization by an enzyme with relatively low affinity for the chain terminating nucleotide leads to exponential amplification whereas an enzyme with relatively high affinity for the chain terminating nucleotide terminates the polymerization and yields sequencing products.

In another aspect, the invention features kits for directly amplifying nucleic acid templates and generating base specifically terminated fragments. In one embodiment, the kit can comprise an appropriate amount of: i) a complete set of chain- elongating nucleotides; ii) at least one chain-terminating nucleotide; (iii) a first DNA polymerase, which has a relatively low affinity towards the chain terminating nucleotide., and (iv) a second DNA polymerase, which has a relatively high affinity towards the chain terminating nucleotide. The kit can also optionally include an appropriate primer or primers, appropriate buffers as well as instructions for use.

The instant invention allows DNA amplification and termination to be performed in one reaction vessel. Due to the use of two polymerases with different affinities for dideoxy nucleotide triphosphates, exponential amplification of the target sequence can be accomplished in combination with a termination reaction nucleotide. In addition, the process obviates the purification procedures, which are required when amplification is performed separately from base terminated fragment generation. Further, the instant process requires less time to accomplish than separate amplification and base specific termination reactions.

When combined with a detection means, the process can be used to detect and/or quantitate a particular nucleic acid sequence where only small amounts of template are available and fast and accurate sequence data acquisition is desirable. For example, when combined with a detection means, the process is useful for sequencing unknown genes or other nucleic acid sequences and for diagnosing or monitoring certain diseases or conditions, such as genetic diseases, chromosomal abnormalities, genetic predispositions to certain diseases (e.g. cancer, obesity, arteriosclerosis) and pathogenic (e.g. bacterial., viral., fungal., protistal) infections. Further, when double stranded DNA molecules are used as the starting material., the instant process

provides an opportunity to simultaneously sequence both strands, thereby providing greater certainty of the sequence data obtained or acquiring sequence information from both ends of a longer template.

1.3.2 Base-specific Reactions Used on DNA fragments from a piece of an unknown sequence

In accordance with the present invention, there is also provided a method and apparatus for determining the sequence of the bases in DNA by measuring the molecular mass of each of the DNA fragments in mixtures prepared by either the Maxam-Gilbert or Sanger-Coulson techniques. The fragments are preferably prepared as in these standard techniques, although the fragments need not be tagged with radioactive tracers. These standard procedures produce from each section of DNA to be sequenced four separate collections of DNA fragments, each set containing fragments terminating at only one or two of the four bases. In the Maxam-Gilbert method, the four separated collections contain fragments terminating at G, both G and A, both C and T, or C positions, respectively. Each of these collections is sequentially loaded into an ultraviolet laser desorption mass spectrometer, and the mass spectrum of each collection is recorded and stored in the memory of a computer. These spectra are recorded under conditions such that essentially no fragmentation occurs in the mass spectrometer, so that the mass of each ion measured corresponds to the molecular weight of one of the DNA fragments in the collection, plus a proton in the positive ion spectrum, and minus a proton in the negative ion spectrum. Spectra obtained from the four spectra are compared using a computer algorithm, and the location of each of the four bases in the sequence is unambiguously determined.

It is also possible, in principle, to obtain the DNA sequence from a single mass spectrum obtained from a more complex single mixture containing all possible fragments, but both the resolution and mass accuracy required are much higher than in the preferred method described above. As a result the accuracy of the DNA sequence obtained from the single spectrum method will generally be inferior, and the gain in raw sequence speed will be counterbalanced by the need for more repetitions to assure accuracy of the sequence.

The DNA fragments to be analyzed are dissolved in a liquid solvent containing a matrix material. Each sample is radiated with a UV laser beam at a wavelength of between 260 nm to 560 nm, and pulses of from 1 to 20 ns pulsewidth.

It is an objective of the present invention to provide a method and apparatus for the rapid and accurate sequencing of human genome and other DNA material.

It is a further objective of the present invention to provide an instrument and method which are relatively simple to operate, relatively low in cost, and which may be automated to sequence thousands of gene bases per hour.

It is a further objective of the present invention to obtain much faster and more accurate DNA sequence data by eliminating the gel electrophoresis separation technique used in conventional DNA sequencing methods to determine the masses of the DNA fragments in a mixture.

1.3.3 Sequencing Through Exposure To Immobilized Probes Of Shorter Length

According to one embodiment of the invention, a target oligonucleotide is exposed to a large number of immobilized probes of shorter length. The probes are collectively referred to as an "array." In the method, one identifies whether a target nucleic acid is complementary to a probe in the array by identifying first a core probe having high affinity to the target, and then evaluating the binding characteristics of all probes with a single base mismatch as compared to the core probe. If the single base mismatch probes exhibit a characteristic binding or affinity pattern, then the core probe is exactly complementary to at least a portion of the target nucleic acid.

The method can be extended to sequence a target nucleic acid larger than any probe in the array by evaluating the binding affinity of probes that can be termed "left" and "right" extensions of the core probe. The correct left and right extensions of the core are those that exhibit the strongest binding affinity and/or a specific hybridization pattern of single base mismatch probes.

The binding affinity characteristics of single base mismatch probes follow a characteristic pattern in which probe/target complexes with mismatches on the 3' or 5' termini are more stable than probe/target complexes with internal mismatches. The process is then repeated to determine additional left and right extensions of the core probe to provide the sequence of a nucleic acid target.

In some embodiments, such as in diagnostics, a target is expected to have a particular sequence. To determine if the target has the expected sequence, an array of probes is synthesized that includes a complementary probe and all or some subset of all single base mismatch probes. Through analysis of the hybridization pattern of the target to such probes, it can be determined if the target has the expected sequence and, if not, the sequence of the target may optionally be determined.

Kits for analysis of nucleic acid targets are also provided by virtue of the present invention. According to one embodiment, a kit includes an array of nucleic acid probes. The probes may include a perfect complement to a target nucleic acid.

The probes also include probes that are single base substitutions of the perfect complement probe. The kit may include one or more of the A, C, T, G, and/or U substitutions of the perfect complement. Such kits will have a variety of uses,

including analysis of targets for a particular genetic sequence, such as in analysis for genetic diseases.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

1.3.4 Sequencing Contiguously, Without The Need For Fragmenting And Sub-Cloning The DNA

The present invention also enables the amplification of a DNA adjacent to a known sequence using the PCR, without the knowledge of the sequence for a second primer.

The present primary invention also provides a new method for sequencing a contiguously very long DNA sequence using the PCR technique, thereby enabling contiguous genomic sequencing. It will avoid the need for mapping or sub-cloning of shorter DNA fragments from haploid genomes such as the bacterial genomes. This method can be used on very large DNA inserts into vectors such as the YAC. Thus, diploid genomes can be sequenced without any further need to sub-clone from the YAC clones. The cloned inserts can be of any length, of several million nucleotides. Alternatively, wherever purified chromosomes are available, this method can be directly applied to sequence the whole chromosome without any need to fragment the chromosome or obtain YAC clones from the chromosome. This method can also be used on whole unpurified genomes with appropriate modifications to account for the allelic variations of the two alleles present on the two chromosomes. In essence, using the method of the present invention, one can generate contiguous genomic sequence information in a manner not possible with any other known protocol using PCR.

The extended invention that enables the sequencing of an unknown region of very long DNA (e.g. genomic DNA) of totally unknown sequence would also find many applications in biology and medicine. For instance, it can be used to physically "map" a chromosome or genome. It would, for example, enable the production of an inventory of many about 500 nucleotide long sequences and the exact primer associated with each of them. This method would also enable the cloning of the amplified DNA sequences from arbitrary regions from a genomic DNA without the need for breaking down the DNA. Using appropriately longer partly fixed primers (as the second primers), very long DNA pieces (several kilobases long) could be amplified and cloned by using this method.

1.3.4.1 PCR Technique with 1 Primer

In one embodiment, the present invention enables the amplification of a DNA stretch using the PCR procedure with the knowledge of only one primer. Using this basic method, the present invention describes a procedure by which a very long DNA

of the order of millions of nucleotides can be sequenced contiguously, without the need for fragmenting and sub-cloning the DNA. In this method, the general PCR technique is used, but the knowledge of only one primer is sufficient, and the knowledge of the other primer is derived from the statistics of the distributions of oligonucleotide sequences of specified lengths.

Present DNA sequencing methods using the separation of DNA fragments on a gel has a limitation of resolving the products of length up to about 1000 nucleotides. Thus, in a single step, the sequence of a DNA fragment up to a length of only about 1000 nucleotides can be obtained by the two conventional DNA sequencing methods. A DNA sequence of a few nucleotides up to many thousand nucleotides can be amplified by the PCR procedure. Thus the PCR procedure can be combined with the DNA sequencing procedure successfully.

A primer is usually of length twelve nucleotides and longer. Let the sequence of one primer is known in a long DNA sequence from which the DNA sequence is to be worked out. From this primer sequence, a specific sequence of four nucleotides occurs statistically at an average distance of 256 nucleotides. It has been worked out by Senapathy that a particular sequence of four characters would occur anywhere from zero distance up to about 1500 characters with a 99.9% probability (P. Senapathy, "Distribution and repetition of sequence elements in eukaryotic DNA: New insights by computer aided statistical analysis," *Molecular Genetics (Life Sciences Advances)*, 7:53-65 (1988)). The mean distance for such an occurrence is 256 characters and the median is 180 characters. Similarly, a 5 nucleotide long specific sequence will occur at a mean distance of 1024 characters, with 99.99% of them occurring within 6000 characters from the first primer. The median distance for the occurrence of a 5-nucleotide specific sequence is ~730 nucleotides. Similarly, a particular 6 nucleotide long sequence will occur at a mean distance of 4096 nucleotides and a median distance of ~2800 nucleotides. A primer of known length, say length 14 can be prepared with a known sequence of 6 characters and the rest of the sequence being random in sequence. It means that any of the four nucleotides can occur at the "random" sequence locations. With a fixed 5, 6 or 7 nucleotide sequence within the second primer, a primer of length 12-18 can be prepared with high specificity of binding.

1.3.4.2 Non-Random Primer (Partly Fixed Primer)

Such a partially non-random primer (hereafter called the partly fixed primer, or partly non-random primer, meaning that part of its sequence is fixed) can "anneal" to only the sequence at which the fixed sequence exists. That is, from the first primer, the partly fixed primer will bind at an average distance of 1024 characters (for a fixed five nucleotide characters). This primer will bind specifically only at the location of the occurrence of the particular five nucleotide sequence with respect to the first primer. The average distance between the first primer and the second non-random primer is ideal for DNA amplification and DNA sequencing. In this situation, the first primer is labeled. Thus, although there would be many locations in the long DNA molecule at which the non-random primer can bind, it would not affect the DNA sequencing because it is dependent only upon the labeled primer.

1.3.4.3 Partly Fixed 2nd Primer

Although the partly fixed second primer has a random sequence component in it, a sub-population of the primer molecules will have the exact sequence that would bind with the exact target sequence. The proportion of the molecules with exact sequence that would bind with the exact target sequence will vary depending on the number of random characters in the partly fixed second primer. For example, in a second primer 11 nucleotides long with 6 characters fixed and 5 characters random, one in ~1000 molecules will have the exact sequence complementary to the target sequence on the template. By increasing the concentration of the partly fixed second primer appropriately, a comfortable level of PCR amplification required for sequencing can be achieved. When primer concentration is increased, it requires an increase in the concentration of Magnesium, which is required for the function of the polymerase enzyme. The excess primers (and "primer- dimers" formed due to excess of primers) can be removed after amplification reaction by a gel-purification step.

Any non-specific binding by any population of the second primers to non-target sequences could be avoided by adjusting (increasing) the temperature of re-annealing appropriately during DNA amplification. It is well known that the change of even one nucleotide due to point-mutation in some cancer genes can be detected by DNA-hybridization. This technique is routinely used for diagnosing particular cancer genes (e.g. John Lyons, "Analysis of ras gene point mutations by PCR and oligonucleotide hybridization," in PCR Protocols: A guide to methods and applications, edited by Michael A Innis et al., (1990), Academic Press, New York).

This is done by adjusting the "re-annealing" or "melting- temperature", and fine-tuning the reaction conditions. Thus the binding of non-specific sequences even with just one nucleotide difference compared to the target binding-site in the template sequence can be avoided.

It should also be noted that non-specific binding sites for the partly fixed second primers could be expected to occur statistically on a long genomic DNA at many places other than the target site which is close to the first primer. Amplification of non-specific DNA between these primer binding sites that could occur on opposite strands of the template DNA could happen. However, this would not affect the objective of the present invention of specific DNA sequencing of the target sequence. Because only the first primer is labeled radioactivity or fluorescently, only the reaction products of the target DNA will be visualized on the sequencing gel pattern. The presence of such non-specific amplification products in the reaction mixture will also not affect the DNA sequencing reaction.

Amplification of DNA will occur not only between the first primer and the partly fixed second primer that occurs closest downstream from the first primer, but also between the first primer and one or two subsequently occurring second primers, depending upon the distance at which they occur. However, these amplification products will all start from the first primer and will proceed up to these second primers. Since the DNA sequencing products are visualized by labeling the first primer, and since the DNA synthesis during the sequencing reaction proceeds from the first primer, the presence of two or three amplification products that start from the first primer will not affect the DNA sequencing products and their visualization on gels. At the most, the intensity of the bands that are subsets of different amplification products will vary slightly on the gel, but not affect the gel pattern. In fact, it is expected that this phenomenon will enable the sequencing of a longer DNA strand where the closest downstream primer is too close to the first primer--thereby avoiding the need for sequencing from the first primer again using another partly fixed second primer.

The minimum length of primer for highly specific amplification between primers on a template DNA is usually considered to be about 15 nucleotides. However, in the present invention, this length can be reduced by increasing the G/C content of the fixed sequence to 12-14 nucleotides.

In essence, the basic procedure of the present invention is fully viable and feasible, and any non-specificity can be avoided by fine-tuning the reaction conditions such as adjusting the annealing temperature and reaction temperature during amplification, and/or adjusting the length and G/C content of the primers, which are routinely done in the standard PCR amplification protocol.

1.3.4.4 Sequence DNA of 2nd Primer

The primary advantage of the present invention is to provide an extremely specific second primer that would bind precisely to a sequence at an appropriate distance from the first primer resulting in the ability to sequence a DNA without the prior knowledge of the second primer. From the newly worked out DNA sequence, a primer sequence can be made complementary to a sequence located close to the downstream end. This can be used as the first primer in the next DNA amplification-sequencing reaction, and the unknown sequence downstream from it can be obtained by again using the same partly fixed primer that was used in the first round of sequencing as the second primer. Thus, knowing only one short sequence in a contiguously long DNA molecule, the entire sequence can be worked out using the present invention.

When the length of the fixed sequence in the partly fixed second primer is increased in the present invention, the distance from the first primer at which the second primer will bind on the template will also be correspondingly increased. For a 6 nucleotide fixed sequence, the median length of DNA amplified will be ~2800 nucleotides (mean 4096 nucleotides), and for a 7 nucleotide fixed sequence, the median length of amplified DNA will be ~11,000 nucleotides (mean ~16,000 nucleotides). However, even if the length of amplified DNA is several thousand nucleotides, still this DNA can be used in DNA sequencing procedures. Furthermore, the present invention can be used to amplify a DNA of length which is limited only by the inherent ability of PCR amplification. A technique known as "long PCR" is used to amplify long DNA sequences (Kainz et al., "In vitro amplification of DNA Fragments > 10 kb," *Anal Biochem.*, 202:46 (1992); Ponce & Micol, "PCR amplification of long DNA fragments" *Nucleic Acids Research*, 20:623 (1992)).

Existing genome sequencing methods employ the breaking down of a very long genomic DNA into many small fragments, sub-cloning them, sequencing them, and then assembling the sequence of the long DNA. Typically, a genomic DNA is

broken down and cloned into overlapping fragments of approx. one million nucleotides in "YAC" (Yeast Artificial Chromosome) clones, each YAC clone is again fragmented and sub-cloned into overlapping fragments of ~25,000 nucleotides in "cosmid" clones, and each cosmid clone in turn sub-cloned into overlapping fragments of ~1000 nucleotides in "M13 phage" or "plasmid" clones. These are sequenced randomly to assemble the larger sequences in the hierarchy. The present invention circumvents the need for breaking down and sub-cloning steps, making it greatly advantageous for contiguously sequencing long genomic DNA.

1.3.4.5 The 2nd Partly Fixed Primer Enabling Sequencing

Extending the above invention, another invention is presented here. This extended invention would enable the sequencing of ~500 nucleotide long sequence somewhere within a given long DNA with no prior information of any sequence at all within the long DNA. The probability that any specific primer of length 10 nucleotides would occur somewhere in a DNA of about one million nucleotides is approximately 1. The probability that any primer of length 15 nucleotides occur somewhere in a genome of about one billion nucleotides is approximately 1. Thus, use of any exact primer of about 15 nucleotide sequence on a genomic DNA in the present invention as the first primer, and the use of the second partly fixed primer will enable the sequencing of the DNA sequence bracketed by the two primers somewhere in the genome. Thus, this procedure can be used to obtain an exact sequence of about 500 characters somewhere from a genome without the prior knowledge of any of its sequence at all. Thus, by using many different primers with arbitrary but exact sequences, one can obtain many ~500-nucleotide sequences at random locations within a genome. Using these sequences as the starting points for contiguous genome sequencing in the present invention, the whole genomic sequence can be closed and completed. Thus an advantage of the present invention is that without any prior knowledge of any sequence in a genome, the whole sequence of a genome can be obtained.

It must be noted that every 15-nucleotide arbitrary primer may not always have a complementary sequence in a genome (of ~one billion nucleotides long). However, most often it would be present and would be useful in performing the above-mentioned sequencing. In some cases, there may be more than one occurrence of the primer sequence in the genome, and so may not be useful in obtaining the

sequence. However, the frequency of successful single-hits can be extremely high (~90%) and can be further refined by using an appropriate length of the arbitrary primer. For genomes (or long DNAs) that are shorter than a billion nucleotides, shorter exact sequences in the first primers (say 10 characters) could be used, and the rest could be random or "degenerate" nucleotides. While this primer will still bind at the sequence complementary to the exact sequence, the longer primer will aid in avoiding non-specific DNA amplification. The length of the first primer can thus be increased using degenerate nucleotides at the ends to a desired extent, without affecting any specificity. Once a sequence is known in an unknown genomic DNA, then the present method can be performed to extend a contiguous sequence in both directions of the DNA from this starting point.

The present invention can also be useful to amplify the DNA between the first primer and the partly fixed second primer, with an aim to using this amplified DNA for purposes other than DNA sequencing, such as cloning. Although there would be sufficient quantity of the target specific amplified DNA in the reaction products, the reaction products will, however, contain the population of non-specific DNA amplified between the non-specifically occurring second primer binding sites on opposite strands. However, by introducing a purification step from this reaction mixture, such as using an immobilized column containing only the first primer, the amplified target DNA can be purified and used for any other purposes.

1.3.5 Sequencing large fragments of DNA (end-sequencing-based method of subclone pathway generation through the fragment with efficient transposon-based sequencing of the identified subclones)

The invention also provides a systematic and efficient way to sequence large fragments of DNA, in particular genomic DNA. It combines an end-sequencing-based method of subclone pathway generation through the fragment with efficient transposon-based sequencing of the identified subclones.

Thus, in one aspect, the invention is directed to a method to sequence a fragment of DNA, said fragment typically having a length of more than about 30 kb. The method comprises the following steps.

First, the fragment is provided in a host cloning vector capable of accommodating it. The size of the fragment that can be sequenced will depend on the nature of the host cloning vector. Cloning vectors are available that can accommodate large fragments of DNA; even the approximately 30-40 kb fragments that are suitable for insertion into cosmids are of sufficient length that the method of the invention is usefully applicable to them.

A composition comprising said vector containing the inserted fragment is then randomly sheared, such as by sonication, to obtain subfragments of approximately 3 kb. The length of the subfragments is appropriate to the transposon-mediated directed sequencing method that will ultimately be applied. The 3 kb length is an approximation; it is intended only as an order of magnitude. Generally speaking, subfragments of 2-5 kb are susceptible to this approach.

The subfragments are then inserted into host cloning vectors to obtain a library of subclones. These host cloning vectors are ideally of minimal size, containing only a selectable marker, an origin of replication, and appropriate insertion sites for the subfragments. The desirability of minimizing the available plasmid DNA in the performance of transposon-mediated sequencing is described by Strathmann, et al. (*supra*).

Sufficient subclones that contain subfragments derived from the original fragment are then recovered to provide 1x coverage of the fragment when the end of each subfragment is sequenced. A stretch of about 400-450 bases can be sequenced with assurance using available automated sequencing techniques. Thus, the sequencing can be conducted using the sequencing primers based on the vector

sequences adjacent the inserts to proceed into the insert to approximately this distance. For a 1 x coverage of the original fragment, the number of subclones required can be calculated by dividing the length of the original fragment by the intended sequencing distance--i.e., by approximately 400- 450.

There should also be sufficient subclones in the library so that when the complete sequence of each is determined, the coverage of the original fragment will be about 7-8 x. This provides, as described below, a high probability that every nucleotide present in the fragment will be present in the library. This number can, of course, be determined by multiplying the length of the fragment by 7 or 8 and dividing by the length of the subfragments generated.

It is preferable to assure that all of the subclones in the library contain pieces of the original fragment. This can be done by recovering only those subclones that hybridize to the fragments.

A sufficient portion of one of the ends of each recovered subclone containing fragment-derived DNA is then sequenced and this sequence information is placed into a searchable database. The database is searched for subclones that contain subfragments with nucleotide sequences matching those that characterize the host vector that accommodated the original fragment. To the extent that these subfragments also contain sequence from the original fragment, that sequence must be at one or the other end of the original fragment. This illustrates why the efficiency of the method is improved by introducing a prescreening step which eliminates any subclones which do not contain portions of the original fragment. If the prescreening has been done, these subclones contain oligonucleotide sequence from either end of the original fragment. The identified subclones are recovered.

1.3.5.1 "Second End" Sequence

A partial sequence of each of the identified subclones is determined from the opposite end of the subfragment insert from that originally placed in the database. This provides "second end" sequence information concerning sequence further removed from the end of the original fragment. This information is then used to search the database in order to identify subclones containing nucleotide sequence that matches this second end sequence. Such subclones are likely to represent regions of the original fragment that are farther removed from the ends and provide further progress in constructing a path across the fragment. These subclones are recovered as

well, and sequenced from the end opposite to that which was sequenced to provide the information for the database and this new information, in turn, used to search the database for a matching sequence. The steps of second end sequencing, searching the database with the resulting sequence information, and recovery of subclones which contain a match are repeated sequentially until subclones have been identified that represent the complete original fragment. The resulting collection of subclones consists of an ordered minimum set that collectively represent the original fragment. The appropriate sequence of such subclones to span the original fragment from end to end is also known.

It remains only to obtain sufficient portions of the complete nucleotide sequence of each subclone from the subclone collection using transposon-mediated sequencing to provide the complete sequence of the original fragment.

In another aspect, the invention is directed to kits suitable for conducting the method of the invention.

1.3.6 Improvements in high speed, high throughput, no required electrophoresis (and, thus, no gel reading artifacts due to the complete absence of an electrophoretic step)

The invention also describes a new method to sequence DNA. The improvements over the existing DNA sequencing technologies include high speed, high throughput, no required electrophoresis (and, thus, no gel reading artifacts due to the complete absence of an electrophoretic, step), and no costly reagents involving various substitutions with stable isotopes. The invention utilizes the Sanger sequencing strategy and assembles the sequence information by analysis of the nested fragments obtained by base-specific chain termination via their different molecular masses using mass spectrometry, for example, MALDI or ES mass spectrometry. A further increase in throughput can be obtained by introducing mass modifications in the oligonucleotide primer, the chain-terminating nucleoside triphosphates and/or the chain- elongating nucleoside triphosphates, as well as using integrated tag sequences which allow multiplexing by hybridization of tag specific probes with mass differentiated molecular weights.

1.3.7 A method and a system for sequencing a genome

The present invention pertains to a method for sequencing genomes. The method comprises the steps of obtaining nucleic acid material from a genome. Then there is the step of constructing a clone library and one or more probe libraries from the nucleic acid material. Next there is the step of comparing the libraries to form comparisons. Then there is the step of combining the comparisons to construct a map of the clones relative to the genome. Next there is the step of determining the sequence of the genome by means of the map.

The present invention pertains to a system for sequencing a genome. The system comprises a mechanism for obtaining nucleic acid material from a genome. The system also comprises a mechanism for constructing a clone library and one or more probe libraries. The constructing mechanism is in communication with the nucleic acid material from a genome. Additionally, the system comprises a mechanism for comparing said libraries to form comparisons. The comparing mechanism is in communication with the said libraries. The system also comprises a mechanism for combining the comparisons to construct a map of the clones relative to the genome. The said combining mechanism is in communication with the comparisons. Further, the system comprises a mechanism for determining the sequence of the genome by means of said map. The said determining mechanism is in communication with said map.

1.3.7.1 A method for producing a gene of a genome

The present invention additionally pertains to a method for producing a gene of a genome. The method comprises the steps of obtaining nucleic acid material from a genome. Then there is the step of constructing libraries from the nucleic acid material. Next there is the step of comparing the libraries to form comparisons. Then there is the step of combining the comparisons to construct a map of the clones relative to the genome. Next there is the step of localizing a gene on the map. Then there is the step of cloning the gene from the map.

1.3.8 Methods and means for the massively parallel characterization of complex molecules and of molecular recognition phenomena with parallelism and redundancy attained through single molecule examination methods

In another embodiment, the present invention approaches the vastness of biological complexity through massive parallelism, which may conveniently be attained through various single molecule examination (SME) methods variously referred to heretofore as single molecule detection (SMD), single molecule visualization (SMV) and single molecule spectroscopy (SMS) techniques.

Used within appropriate procedures, single molecule examination methods can enable molecular parallelism.

Molecular parallelism may be applied to the examination of the composition of complex molecules (including co-polymers of natural or of synthetic origin) or to determinations of interactions between large numbers of molecules. The former case may be applied to genome-scale sequencing methods. The latter case may be applied to rapid determination of molecular complementarity, with applications in (biological or non-biological) affinity characterization, immunological study, clinical pathology, molecular evolution (e.g. in vitro evolution), and the construction of a cybernetic immune system as well as prostheses based thereupon. In both cases, molecular recognition phenomena are observed with molecular parallelism.

Note that within said affinity characterization applications, both kinetics of both binding association and dissociation, and binding equilibria, may be examined. Kinetics may be examined by observing the rates of occupation of appropriate sites or diverse populations thereof by some homogenous or heterogeneous sample, and the rates of vacancy formation from occupied sites. Equilibria constants may be determined by observing the proportion (number of occupied sites divided by number of total sites) of sites occupied under equilibrium conditions, with greater quantitative confidence yielded by, for example, examining more binding sites.

Sequencing of polynucleotide molecules may be effected by the (preferably end-wise) immobilization of a library of such molecules to a surface at a density convenient for detection, which will vary according to the detection methodology availed. Several methods capable of effecting such immobilization will be obvious to those skilled in the arts of recombinant DNA technology and molecular biology,

among others. Priming, which may be random or non-random, is effected by any of a variety of methods, most of which are obvious to those skilled in the relevant arts.

Genome sequencing applications availing of enzymatic polymerization's and corresponding embodiments of the present invention, rely upon control over polymerization rate and nucleotide incorporation specificity, consistent with the well-known Watson-Crick base pairing rules which may be enforced (upon single nucleotides in a processive manner, as conditions permit) by the use of DNA polymerases or analogs thereof, in combination with repeatable single molecule detection applied to a large population of diverse molecules. A sequencing cycle comprises the steps of: (1.) polymerizing one or less nucleotides, which carry some removable or neutralizable molecular label and may optionally be reversibly 3' protected (or otherwise protected in any manner which modulates polymerization rate onto each sample molecule at the primer or at subsequent extensions thereof and in opposition to (and pairing with) a single, unique, base of the template polynucleotide strand; (2.) optionally washing away any unreacted labeled nucleotides; (3.) detecting, by either direct or indirect methods, said labeled nucleotides incorporated into said sample molecules, in a manner which repeatably associates information obtained about the type of label observed with the unique identity of the template molecule under observation, which may be uniquely distinguished by a variety of methods (which include: a mappable location of immobilization of the sample template molecule on a substrate surface; a mappable location of immobilization of the sample template molecule within some matrix volume element; microscopic labeling with some readily identifiable, e.g. combinatorially or permutationally diverse and readily examined particle or molecule or group of molecules and detection of the thus marked identity of individual free molecules in solution; and, scanning of a liquid sample

may serve to modulate monomer addition rate to the strand being copied from the template molecule) from the nucleotide added during the present cycle, if these are distinct from any cleavably linked labeling moieties; (6.) optionally checking that the removal or neutralization of said label in step (4) was successful for any particular molecule of the sample, by repeating a similar detection procedure. Said sequencing cycle comprising an appropriate subset of steps 1-6 may be repeated as many times as convenient, but must be repeated a sufficient number of times to obtain sequence information of sufficient complexity from each individual molecule to permit unambiguous alignment of all such sequence information determined for all of the molecules of the sample. This minimum number of cycles will be approximately related to the complexity C of the sample to be treated as part of the same macroscopic reaction (i.e. a macroscopic sample preparation subjected to unitary macroscopic manipulations) by the formula $C < 4^n$ where n is the number of cycles. Beyond this minimum, there are tradeoffs between the number of cycles to be performed and the number of molecules to be examined, and the confidence for sequence data obtained.

Note that unused reagents and enzymes may be recovered from washes and recycled.

1.3.8.1 Advantages of Parallelism

In contrast to the previously disclosed base-addition sequencing schemes, the sequence determination applications of the present invention enjoys substantial advantages deriving from sample manipulation in the single-molecule-regime. Working instead in the distinct single-molecule-regime rather than with populations of identical molecules provides substantial advantages of parallelism, facility of use and implementation, (including automated implementation,) and operability. Among these are unanticipated advantages: (1) because a single molecule is necessarily monodisperse, failure of a molecule to undergo addition in a cycle does not cause a loss of sample monodispersion (i.e. lead to uneven sample molecules dispersity or polydispersion); such addition failure is unproblematic when single molecules are examined individually because it is readily detected and accounted for in data analysis; in contrast, samples comprising multiple identical molecules may thus take on non-identical lengths, complicating data collection and analysis; (2) samples comprising a plurality of individually distinct single molecules (species) may be

handled unitarily without requiring any handling measures to keep distinct molecules apart, providing a large reduction in manipulations required on a per-species basis and not requiring the use of many separate, parallel fluid handling steps or means; (3) inadvertent multiple base additions are more readily detected and their extent is more readily quantified because these changes in quantity are large compared to the signal expected from the incorporation of a single base (i.e. single label) into a single molecular species; (4) deprotection or delabeling failures may also be readily detected and noted for the correct single molecule, such that addition failure, the presence of a label, or overlabeling in the subsequent cycle may be correctly interpreted (according to the unlabeled and single stepping methods used in a particular embodiment.) These advantages are expected to be important in the competitiveness of these present methods over conventional polynucleotide sequencing methods.

Various techniques are included to address any non-idealities encountered which may arise because of deviations from conventional polymerization or detection methods. These generally take the form of different types of redundancy, which may be employed to either prevent or resolve any such errors. Prominent among these redundancies is oversampling, i.e. the examination of some multiple (j) of the number (m) of sample molecules suggested by combinatoric computations to be minimally sufficient for full alignment of data from a sample of a given complexity.

Such oversampling redundancy will increase the confidence interval for accuracy of collected data and reduce the likelihood of artifacts arising from sequence duplications which may occur in any given sample.

1.3.8.2 Oversampling Redundancy

Oversampling redundancy may be availed to increase data confidence by providing the opportunity to score and match multiple occurrences of the same sequence segment and thus detect and eliminate erroneous sequence segment information by virtue of its less frequent occurrence. Erroneous sequence segment information may arise, for instance, by nucleotide incorporation errors which are an inevitable feature of polymerization with polymerases having a characteristic fidelity, i.e. displaying a characteristic nucleotide misincorporation rate. Such methods will be particularly useful where polynucleotide polymerases fidelity would otherwise be unacceptably low. It should be noted that an error rate of one percent or more has been deemed conventionally acceptable for genome informatic purposes.

1.3.8.3 Controls/Data

Further, known molecules having sequences that are highly unrelated to the sample may be included as internal controls to monitor the efficiency and accuracy of a particular sequence collection process; such internal control sequences will present negligibly small overhead because molecular parallelism may easily accommodate any such comparatively small increase in sample complexity, even though it might be considered large with respect to pre-existing methods.

After raw data have been collected for each molecule, these are all mutually compared by some appropriate matching algorithm and aligned so as to reconstruct the full sequence of the sample. The computational complexity of completing such an alignment may be estimated as the multi-phase comparison and sorting of $(j)(m)$ strings each of length n .

Alternatively, data alignment may be performed in tandem or parallel with later cycles and may be monitored by appropriate computational algorithms for data quality and confidence of sequence information, and cycling may continue till desired criteria are satisfied. Computer, microprocessor, electronic or other automated control of instrumentation, including fluidics and robotics for the manipulation of samples, and the automated effectuation of the various methods of the present invention, all according to parameterized algorithms, may be accomplished by means obvious from the present disclosure to those skilled in the relevant arts (e.g. fluidics, robotics, electronics, microelectronics, computer science and engineering, and mechanical engineering). Concurrent data alignment and monitoring will permit modifications of the sequencing cycle described above, such as dynamic adjustment of polymerization reaction conditions and durations, label removal or neutralization procedure parameters, polymerization deprotection conditions, and any other desired parameter, so as to permit optimization of procedures and results.

With appropriately flexible design, automated systems and instruments such as those described above for genome applications may readily be adapted, with appropriate changes in samples and labeling methods and reagents, to cybernetic molecular evolution, cybernetic immune system, broad spectrum pathogen characterization and other applications of the present invention.

1.3.8.4 Double/Single Stranded Polynucleotide Sequencing Method

According to the embodiment availed, double or single stranded polynucleotides may be examined. Where single stranded polynucleotide molecules are preferred, second strands may be removed by performing said immobilization so as to only involve only one strand in covalent linkage with said surface and then performing a denaturation of the sample with washing. Priming means required by any particular enzyme must then be provided, usually by hybridization of a complementary oligo- or polynucleotide to the sample template molecules, though other means are possible. Other methods which will be obvious to those skilled in the arts of recombinant DNA technology may also be employed to yield immobilized or otherwise uniquely identifiable single stranded polynucleotide samples.

Where double stranded molecules are preferred, said second strands may be treated with an appropriate exonuclease under appropriate conditions and for an appropriate lengths of time to provide a good distribution of lengths of said second strands such that the termini of the undegraded portions of said second strands provide convenient priming for enzymatic nucleotide polymerization (i.e. DNA directed DNA synthesis or DNA replication, DNA directed RNA synthesis or transcription, RNA directed DNA synthesis or reverse transcription, or RNA directed RNA synthesis or RNA replication).

Note that the polynucleotide sequencing methods of the present invention represent the converse of conventional enzymatic and chemical sequencing methods in that those conventional methods rely upon the production of multiple homogeneous sub-populations of DNA molecules which together comprise a nested set, and the detection of each of such sub-population (with deviant chain terminator misincorporation molecules arising with significantly lower frequency and thus constituting a poorly detected population), while the present invention relies on alignment of information from a highly inhomogeneous population molecules and repeatable detection of single molecules. Further note that by previous methods, each species yields information about only one base at one position within the sample sequence, while with the methods of the present invention, each individual sample template molecule may yield information about the identity of several bases. Note also that under conventional methods, some effort has been expended in increasing the number of bases yielding information per sample, i.e. lengthening the linear sequence information obtained from any one segment of a sample, which is

substantially frustrated by the inherent limitations of electrophoretic separation and particularly gel electrophoresis, while the present invention readily accomplishes the information yielded per unitary manipulation through increases in the facility and practicable extent of parallelism.

There are several levels of parallelism and pipelining possible with the methods of the present invention. An arbitrarily large number of molecules may be subjected to any given manipulation at once if they are part of the same unitary sample. Detection will have constraints entailed by the particular instrumentation and method used, but many degrees of freedom exist with regard to means of providing parallelism in detection instrumentation (e.g. multiple microscopy instruments or appropriately arranged objective lenses and controlled light paths for light microscopic based detection, multiple optoelectronic device arrays [e.g. CCDs or SLMs] for the respective types of detection; multiple probes [i.e. in arrays with parallel detection provided] for scanning probe microscopic detection methods with various degrees of freedom with respect to each other during scanning, etc.) Means for pipelining the steps of the methods disclosed herein will be readily apparent when one considers that dedicated instrumentation or robotics may perform each relevant step, and that the ensemble of such instrumentation may readily be integrated to form a coordinated system, for example matching throughput at different stages by adjusting the parallelism of appropriate stages. Thus economy, throughput and data accuracy are tradeoffs, but may individually vastly exceed any such measures attainable with conventional methods.

1.4 EXEMPLARY GENOMIC CHARACTERIZATION METHODS

1.4.1 Employing Mass Spectrometry To Analyze The Sanger Sequencing Reaction Mixtures

In one embodiment, this invention describes an improved method of sequencing DNA. In particular, this invention employs mass spectrometry to analyze the Sanger sequencing reaction mixtures.

In Sanger sequencing, four families of chain-terminated fragments are obtained. The mass difference per nucleotide addition is 289.19 for dpC, 313.21 for dpA, 329.21 for dpG and 304.2 for dpT, respectively.

1.4.1.1 Mass Modified

In one embodiment, through the separate determination of the molecular weights of the four base-specifically terminated fragment families, the DNA sequence can be assigned via superposition (e.g., interpolation) of the molecular weight peaks of the four individual experiments. In another embodiment, the molecular weights of the four specifically terminated fragment families can be determined simultaneously by MS, either by mixing the products of all four reactions run in at least two separate reaction vessels (i.e., all run separately, or two together, or three together) or by running one reaction having all four chain-terminating nucleotides (e.g., a reaction mixture comprising dTTP, ddTTP, dATP, ddATP, dCTP, ddCTP, dGTP, ddGTP) in one reaction vessel. By simultaneously analyzing all four base-specifically terminated reaction products, the molecular weight values have been, in effect, interpolated. Comparison of the mass difference measured between fragments with the known masses of each chain-terminating nucleotide allows the assignment of sequence to be carried out. In some instances, it may be desirable to mass modify, as discussed below, the chain-terminating nucleotides so as to expand the difference in molecular weight between each nucleotide. It will be apparent to those skilled in the art when mass-modification of the chain-terminating nucleotides is desirable and can depend, for instance, on the resolving ability of the particular spectrometer employed. By way of example, it may be desirable to produce four chain-1 2 3 1 terminating nucleotides, ddTTP, ddCTP, ddATP and ddGTP where ddCTP ddATP 2 and ddGTP 3 have each been mass-modified so as to have molecular weights resolvable from one another by the particular spectrometer being used.

The terms chain-elongating nucleotides and chain-terminating nucleotides are well known in the art. For DNA, chain-elongating nucleotides include 2'-deoxyribonucleotides and chain-terminating nucleotides include 2', 3'-dideoxyribonucleotides. For RNA, chain-elongating nucleotides include ribonucleotides and chain-terminating nucleotides include 3'-deoxyribonucleotides. The term nucleotide is also well known in the art. For the purposes of this invention, nucleotides include nucleoside mono-, di-, and triphosphates. Nucleotides also include modified nucleotides such as phosphorothioate nucleotides.

Since mass spectrometry is a serial method, in contrast to currently used slab gel electrophoresis which allows several samples to be processed in parallel, in another embodiment of this invention, a further improvement can be achieved by multiplex mass spectrometric DNA sequencing to allow simultaneous sequencing of more than one DNA or RNA fragment. As described in more detail below, the range of about 300 mass units between one nucleotide addition can be utilized by employing either mass modified nucleic acid sequencing primers or chain-elongating and/or terminating nucleoside triphosphates so as to shift the molecular weight of the base-specifically terminated fragments of a particular DNA or RNA species being sequenced in a predetermined manner. For the first time, several sequencing reactions can be mass spectrometrically analyzed in parallel. In yet another embodiment of this invention, multiplex mass spectrometric DNA sequencing can be performed by mass modifying the fragment families through specific oligonucleotides (tag probes) which hybridize to specific tag sequences within each of the fragment families. In another embodiment, the tag probe can be covalently attached to the individual and specific tag sequence prior to mass spectrometry.

1.4.1.2 Mass Spectrometer Formats Used (MALDI, ES, ICR, Fourier Transform)

Preferred mass spectrometer formats for use in the invention are matrix assisted laser desorption ionization (MALDI), electrospray (ES), ion cyclotron resonance (ICR) and Fourier Transform. For ES, the samples, dissolved in water or in a volatile buffer, are injected either continuously or discontinuously into an atmospheric pressure ionization interface (API) and then mass analyzed by a quadrupole. The generation of multiple ion peaks which can be obtained using ES mass spectrometry can increase the accuracy of the mass determination. Even more

detailed information on the specific structure can be obtained using an MS/N4S quadrupole configuration. In MALDI mass spectrometry, various mass analyzers can be used, e.g., magnetic sector/magnetic deflection instruments in single or triple quadrupole mode (MS/MS), Fourier transform and time-of-flight (TOF) configurations as is known in the art of mass spectrometry. For the desorption/ionization process, numerous matrix/laser combinations can be used. Ion-trap and reflectron configurations can also be employed.

In one embodiment of the invention, the molecular weight values of at least two base-specifically terminated fragments are determined concurrently using mass spectrometry. The molecular weight values of preferably at least five and more preferably at least ten base-specifically terminated fragments are determined by mass spectrometry. Also included in the invention are determinations of the molecular weight values of at least 20 base-specifically terminated fragments and at least 30 base-specifically terminated fragments. Further, the nested base-specifically terminated fragments in a specific set can be purified of all reactants and by-products but are not separated from one another. The entire set of nested base-specifically terminated fragments is analyzed concurrently and the molecular weight values are determined. At least two base-specifically terminated fragments are analyzed concurrently by mass spectrometry when the fragments are contained in the same sample.

1.4.1.3 Process of Mass Spectrometric DNA Sequencing

In general, the overall mass spectrometric DNA sequencing process will start with a library of small genomic fragments obtained after first randomly or specifically cutting the genomic DNA into large pieces which then, in several subcloning steps, are reduced in size and inserted into vectors like derivatives of M 13 or pUC (e.g., M13mpl8 or M13mpl9). In a different approach, the fragments inserted in vectors, such as M 13, are obtained via subcloning starting with a cDNA library. In yet another approach, the DNA fragments to be sequenced are generated by the polymerase chain reaction (e.g., Higuchi et al., "A General Method of in vitro Preparation and Mutagenesis of DNA Fragments: Study of Protein and DNA Interactions," *Nucleic Acids Res.*, 16, 7351-67 (1988)). As is known in the art, Sanger sequencing can start from one nucleic acid primer (UP) binding to the plus-strand or from another nucleic acid primer binding to the opposite minus-strand. Thus, either

the complementary sequence of both strands of a given unknown DNA sequence can be obtained (providing for reduction of ambiguity in the sequence determination) or the length of the sequence information obtainable from one clone can be extended by generating sequence information from both ends of the unknown vector- inserted DNA fragment.

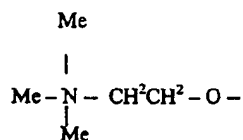
The nucleic acid primer carries, preferentially at the 5'-end, a linking functionality, L, which can include a spacer of sufficient length and which can interact with a suitable functionality, L', on a solid support to form a reversible linkage such as a photocleavable bond. Since each of the four Sanger sequencing families starts with a nucleic acid primer this fragment family can be bound to the solid support by reacting with functional groups, L', on the surface of a solid support and then intensively washed to remove all buffer salts, triphosphates, enzymes, reaction by- products, etc. Furthermore, for mass spectrometric analysis, it can be of importance at this stage to exchange the cation at the phosphate backbone of the DNA fragments in order to eliminate peak broadening due to a heterogeneity in the cations bound per nucleotideunit. Since the L-L' linkage is only of a temporary nature with the purpose to capture the nested Sanger DNA or RNA fragments to properly condition them for mass spectrometric analysis, there are different chemistries which can serve this purpose. In addition to the examples given in which the nested fragments are coupled covalently to the solid support, washed, and cleaved off the support for mass spectrometric analysis, the temporary linkage can be such that it is cleaved under the conditions of mass spectrometry, i.e., a photocleavable bond such as a charge transfer complex or a stable organic radical. Furthermore, the linkage can be formed with L'being a quaternary ammonium group. In this case, preferably, the surface of the solid support carries negative charges which repel the negatively charged nucleic acid backbone and thus facilitates desorption. Desorption will take place either by the heat created by the laser pulse and/or, depending on L', by specific absorption of laser energy which is in resonance with the L' chromophore. The functionalities, L and L', can also form a charge transfer complex and thereby form the temporary L-L' linkage. Various examples for appropriate functionalities with either acceptor or donator properties are depicted without limitation herein. Since in many cases the "charge-transfer band" can be determined by UV/vis spectrometry (see e.g. Organic Charge Transfer Complexes by R. Foster, Academic Press, 1969),

the laser energy can be tuned to the corresponding energy of the charge-transfer wavelength and, thus, a specific desorption off the solid support can be initiated. Those skilled in the art will recognize that several combinations can serve this purpose and that the donor functionality can be either on the solid support or coupled to the nested Sanger DNA/RNA fragments or vice versa.

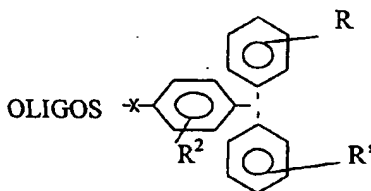
In yet another approach, the temporary linkage L-L' can be generated by homolytically forming relatively stable radicals. As described herein, a combination of the approaches using charge-transfer complexes and stable organic radicals is shown. Here, the nested Sanger DNA/RNA fragments are captured via the formation of a charge transfer complex. Under the influence of the laser pulse, desorption (as discussed above) as well as ionization will take place at the radical position. In other examples described herein, under the influence of the laser pulse, the L-L' linkage will be cleaved and the nested Sanger DNA/RNA fragments desorbed and subsequently ionized at the radical position formed. Those skilled in the art will recognize that other organic radicals can be selected and that, in relation to the dissociation energies needed to homolytically cleave the bond between them, a corresponding laser wavelength can be selected (see e.g. *Reactive Molecules* by C. Wentrup, John Wiley & Sons, 1984). In yet another approach, the nested Sanger DNA/RNA fragments are captured via Watson-Crick base pairing to a solid support- bound oligonucleotide complementary to either the sequence of the nucleic acid primer or the tag oligonucleotide sequence. The duplex formed will be cleaved under the influence of the laser pulse and desorption can be initiated. The solid support- bound base sequence can be presented through natural oligoribo- or oligodeoxyribonucleotide as well as analogs (e.g. thio-modified phosphodiester or phosphotriester backbone) or employing oligonucleotide mimetics such as PNA analogs (see e.g. Nielsen et al., *Science*, 254, 1497 (1991)) which render the base sequence less susceptible to enzymatic degradation and hence increases overall stability of the solid support-bound capture base sequence. With appropriate bonds, L-L', a cleavage can be obtained directly with a laser tuned to the energy necessary for bond cleavage. Thus, the immobilized nested Sanger fragments can be directly ablated during mass spectrometric analysis.

1.4.1.3.1 Conditioning

Prior to mass spectrometric analysis, it may be useful to "condition" nucleic acid molecules, for example to decrease the laser energy required for volatilization, to minimize fragmentation or to otherwise increase the sensitivity of mass spectrometric detection. For example, nucleic acids can be "conditioned" by adding positive or negative charges, i.e. charge tags (CTs). CTs increase the mass spectrometer detection sensitivity by increasing the degree of ionization during the mass spectrometric (e.g. MALDI) process. A CT can be linked either to the external 3' or 5' position or internally e.g. at the 2' position or at the base, e.g. at C-5 in uracil, C-5 methyl group of thymine, C-5 at cytosine, at C⁷ or C⁸ guanine, adenine and hypoxanthine or at the phosphate ester moiety. Charge tags, CTs, can function molecules with permanent (i.e. pH-independent) ionization, such as:



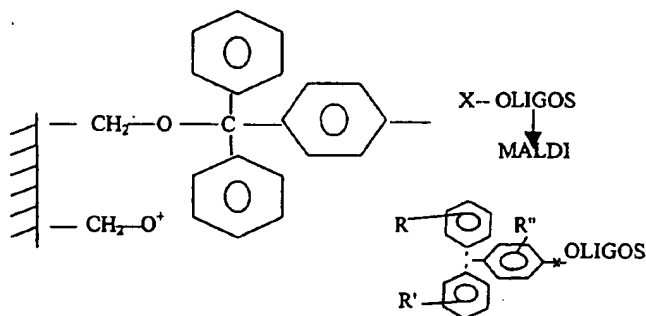
or molecules which generate a positive charge upon MALDI and which are stabilized by delocalization of the positive charge by mesomeric effects in unsaturated and/or aromatic systems such as:



wherein, R, R', R' = H, or Al (wherein Al = e.g. lower alkyl, methyl, ethyl, propyl), NO₂, CN, CO₂H, CO₂ active ester, or halogen; and X = -O-, -NH-, -S-, C=O, OCO either in the para or meta position.

For example, the positive charge of a trityl cation is produced during MALDI by the removal of a moiety such as: -OR, where R = a lower alkyl, or an anion such as ClO_4^- , SbF_6^- , BF_4^- and the like.

In an alternative scheme, the trityl group is used to anchor the oligonucleotide to a solid support via the tertiary carbon and this bond is cleaved during mass spectrometry (e.g. MALDI), leaving a positive charge on the desorbing and high vacuum flying oligonucleotide.



One of skill in the art can readily appreciate several variations to the schemes described above. In addition to employing the charge tag array alone, one of skill in the art can employ a charge tag array in conjunction with another conditioning means. Particularly preferred means to be used in conjunction with the CT include treating the phosphodiester bond with trialkylsilyl halides or the phosphomonothiodiester bond with alkyl iodides to render the polyanionic backbone neutral.

1.4.1.3.1.1 Modification of phosphodiester Backbone of Nucleic Acid Molecule

Another example of conditioning is modification of the phosphodiester backbone of the nucleic acid molecule (e.g. cation exchange), which can be useful for eliminating peak broadening due to a heterogeneity in the cations bound per nucleotide unit. In addition, a nucleic acid molecule can be contacted with an alkylating agent such as alkyl iodide, iodoacetamide, β -iodoethanol, or 2,3-epoxy-1-propanol, the monothio phosphodiester bonds of a nucleic acid molecule can be transformed into a phosphotriester bond. Likewise, phosphodiester bonds may be transformed to uncharged derivatives employing trialkylsilyl chlorides. Further conditioning involves incorporating nucleotides which reduce sensitivity for depurination (fragmentation during MS) such as N7- or N9-deazapurine nucleotides, or RNA building blocks or using oligonucleotide triesters or incorporating phosphorothioate functions which are alkylated or employing oligonucleotide mimetics such as PNA.

Modification of the phosphodiester backbone can be accomplished by, for example, using alpha-thio modified nucleotides for chain elongation and termination. With alkylating agents such as alkyl iodides, iodoacetamide, β -iodoethanol, 2,3-epoxy-1-propanol, the monothio phosphodiester bonds of the nested Sanger fragments are transformed into phosphotriester bonds. Multiplexing by mass modification in this case is obtained by mass-modifying the nucleic acid primer (UP) or the nucleoside triphosphates at the sugar or the base moiety. To those skilled in the art, other modifications of the nested Sanger fragments can be envisioned. In one embodiment of the invention, the linking chemistry allows one to cleave off the so-purified nested DNA enzymatically, chemically or physically. By way of example, the L-L' chemistry can be of a type of disulfide bond (chemically cleavable, for example, by mercaptoethanol or dithioerythrol), a biotin/streptavidin system, a heterobifunctional derivative of a trityl ether group (Koster et al., "A Versatile Acid-Labile Linker for Modification of Synthetic Biomolecules," *Tetrahedron Letters* 31, 7095 (1990)) which can be cleaved under mildly acidic conditions, a levulinyl group cleavable under almost neutral conditions with a hydrazinium/acetate buffer, an arginine-arginine or lysine-lysine bond cleavable by an endopeptidase enzyme like trypsin or a pyrophosphate bond cleavable by a pyrophosphatase, a photocleavable bond which can be, for example, physically cleaved and the like. Optionally, another cation exchange can be performed prior to mass spectrometric analysis. In the instance that an enzyme-cleavable bond is utilized to immobilize the nested fragments, the enzyme used to cleave the bond can serve as an internal mass standard during MS analysis.

1.4.1.3.2 Purification Process

The purification process and/or ion exchange process can be carried out by a number of other methods instead of, or in conjunction with, immobilization on a solid support. For example, the base-specifically terminated products can be separated from the reactants by dialysis, filtration (including ultrafiltration), and chromatography.

Likewise, these techniques can be used to exchange the cation of the phosphate backbone with a counter-ion which reduces peak broadening.

The base-specifically terminated fragment families can be generated by standard Sanger sequencing using the Large Klenow fragment of *E. coli* DNA polymerase I, by Sequenase, Taq DNA polymerase and other DNA polymerases

suitable for this purpose, thus generating nested DNA fragments for the mass spectrometric analysis. It is, however, part of this invention that base-specifically terminated RNA transcripts of the DNA fragments to be sequenced can also be utilized for mass spectrometric sequence determination. In this case, various RNA polymerases such as the SP6 or the T7 RNA polymerase can be used on appropriate vectors containing, for example, the SP6 or the T7 promoters (e.g. Axelrod et al., "Transcription from Bacteriophage T7 and SP6 RNA Polymerase Promoters in the Presence of 3' Deoxyribonucleoside 5' triphosphate Chain Terminators," *Biochemistry* 24, 5716-23 (1985)). In this case, the unknown DNA sequence fragments are inserted downstream from such promoters. Transcription can also be initiated by a nucleic acid primer (Pitulle et al., "Initiator Oligonucleotides for the Combination of Chemical and Enzymatic RNA Synthesis," *Gene* 112, 101-105 (1992)) which carries, as one embodiment of this invention, appropriate linking functionalities, L, which allow the immobilization of the nested RNA fragments, as outlined above, prior to mass spectrometric analysis for purification and/or appropriate modification and/or conditioning.

1.4.1.3.3 Immobilization Process

For this immobilization process of the DNA/RNA sequencing products for mass spectrometric analysis, various solid supports can be used, e.g., beads (silica gel, controlled pore glass, magnetic beads, Sephadex/Sepharose beads, cellulose beads, etc.), capillaries, glass fiber filters, glass surfaces, metal surfaces or plastic material. Examples of useful plastic materials include membranes in filter or microtiter plate formats, the latter allowing the automation of the purification process by employing microtiter plates which, as one embodiment of the invention, carry a permeable membrane in the bottom of the well functionalized with L'. Membranes can be based on polyethylene, polypropylene, polyamide, polyvinylidenedifluoride and the like. Examples of suitable metal surfaces include steel, gold, silver, aluminum, and copper. After purification, cation exchange, and/or modification of the phosphodiester backbone of the L-L' bound nested Sanger fragments, they can be cleaved off the solid support chemically, enzymatically or physically. Also, the L-L' bound fragments can be cleaved from the support when they are subjected to mass spectrometric analysis by using appropriately chosen L-L linkages and corresponding laser energies/intensities as described above and herein.

1.4.1.4 Data Analysis (ES, MALDI)

The highly purified, four base-specifically terminated DNA or RNA fragment families are then analyzed with regard to their fragment lengths via determination of their respective molecular weights by MALDI or ES mass spectrometry.

For ES, the samples, dissolved in water or in a volatile buffer, are injected either continuously or discontinuously into an atmospheric pressure ionization interface (API) and then mass analyzed by a quadrupole. With the aid of a computer program, the molecular weight peaks are searched for the known molecular weight of the nucleic acid primer (UP) and determined which of the four chain terminating nucleotides has been added to the UP. This represents the first nucleotide of the unknown sequence. Then, the second, the third, the n^{th} extension product can be identified in a similar manner and, by this, the nucleotide sequence is assigned. The generation of multiple ion peaks which can be obtained using ES mass spectrometry can increase the accuracy of the mass determination.

1.4.1.5 Process for Multiplex Mass Spectrometric DNA Sequencing Employing Mass Modified Reagents

As illustrative embodiments of this invention, three different basic processes for multiplex mass spectrometric DNA sequencing employing the described mass-modified reagents are described below:

A) Multiplexing by the use of mass-modified nucleic acid primers (UP) for Sanger DNA or RNA sequencing,

B) Multiplexing by the use of mass-modified nucleoside triphosphates as chain elongators and/or chain terminators for Sanger DNA or RNA sequencing, and

C) Multiplexing by the use of tag probes which specifically hybridize to tag sequences which are integrated into part of the four Sanger DNA/RNA base-specifically terminated fragment families. Mass modification here can be achieved as described hereing, or alternately, by designing different oligonucleotide sequences having the same or different length with unmodified nucleotides which, in a predetermined way, generate appropriately differentiated molecular weights.

The process of multiplexing by mass-modified nucleic acid primers (UP) is illustrated by way of example herein for mass analyzing four different DNA clones simultaneously. The first reaction mixture is obtained by standard Sanger DNA

sequencing having unknown DNA fragment 1 (clone 1) integrated in an appropriate vector (e.g., M13mpl8), employing an unmodified nucleic acid primer UP^0 , and a standard mixture of the four unmodified deoxynucleoside triphosphates, $dNTP^0$ and with 1/10th of one of the four dideoxynucleoside triphosphates, $ddNTP$. A second reaction mixture for DNA fragment 2 (clone 2) is obtained by employing a mass-modified nucleic acid primer UP^1 and, as before, the four unmodified nucleoside triphosphates, $dNTP$, containing in each separate Sanger reaction 1/10th of the chain-terminating unmodified dideoxynucleoside triphosphates $ddNTP$. In the other two experiments, the four Sanger reactions have the following compositions: DNA fragment 3 (clone 3), UP^2 , $dNTP^0$, $ddNTP^0$ and DNA fragment 4 (clone 4), UP^3 , $dNTP^0$, $ddNTP^0$. For mass spectrometric DNA sequencing, all base-specifically terminated reactions of the four clones are pooled and mass analyzed. The various mass peaks belonging to the four dideoxy-terminated (e.g., ddT-terminated) fragment families are assigned to specifically elongated and ddT-terminated fragments by searching (such as by a computer program) for the known molecular ion peaks of UP^0 , UP^1 , UP^2 and UP^3 extended by either one of the four dideoxynucleoside triphosphates, $UP^0 ddN^0$, $UP^1 ddN^0$, $UP^2 ddN^0$ and $UP^3 - ddN^0$. In this way, the first nucleotides of the four unknown DNA sequences of clone 1 to 4 are determined. The process is repeated, having memorized the molecular masses of the four specific first extension products, until the four sequences are assigned. Unambiguous mass/sequence assignments are possible even in the worst case scenario in which the four mass-modified nucleic acid primers are extended by the same dideoxynucleoside triphosphate, the extension products then being, for example, $UP^0 ddT$, $UP^1 -ddT$, $UP^2 -ddT$ and $UP^3 -ddT$, which differ by the known mass increment differentiating the four nucleic acid primers. In another embodiment of this invention, an analogous technique is employed using different vectors containing, for example, the SP6 and/or T7 promoter sequences, and performing transcription with the nucleic acid primers UP^0 , UP^1 , UP^2 and UP^3 and either an RNA polymerase (e.g., SP6 or T7 RNA polymerase) with chain-elongating and terminating unmodified nucleoside triphosphates NTP^0 and 3'- $dNTP^0$. Here, the DNA sequence is being determined by Sanger RNA sequencing.

Illustrated herein is the process of multiplexing by mass-modified chain-elongating or/and terminating nucleoside triphosphates in which three different DNA

fragments (3 clones) are mass analyzed simultaneously. The first DNA Sanger sequencing reaction (DNA fragment 1, clone 1) is the standard mixture employing unmodified nucleic acid primer UP^0 , $dNTP^0$ and in each of the four reactions one of the four $ddNTP^0$. The second (DNA fragment 2, clone 2) and the third (DNA fragment 3, clone 3) have the following contents: UP^0 , $dNTP^0$, $ddNTP^1$ and UP^0 , $dNTP^0$, $ddNTP^2$, respectively. In a variation of this process, an amplification of the mass increment in mass-modifying the extended DNA fragments can be achieved by either using an equally mass-modified deoxynucleoside triphosphate (i.e., $dNTP^1$, $dNTP^2$) for chain elongation alone or in conjunction with the homologous equally mass-modified dideoxynucleoside triphosphate. For the three clones depicted above, the contents of the reaction mixtures can be as follows: either $UP^0/dNTP^0/ddNTP^0$, $UP^0/dNTP^1/ddNTP^0$ and $UP^0/dNTP^2/ddNTP^0$ or $UP^0/dNTP^0/ddNTP^0$, $UP^0/dNTP^1/ddNTP^1$ and $UP^0/dNTP^2/ddNTP^2$. As described above, DNA sequencing can be performed by Sanger RNA sequencing employing unmodified nucleic acid primers, UP , and an appropriate mixture of chain-elongating and terminating nucleoside triphosphates. The mass-modification can be again either in the chain-terminating nucleoside triphosphate alone or in conjunction with mass-modified chain-elongating nucleoside triphosphates. Multiplexing is achieved by pooling the three base-specifically terminated sequencing reactions (e.g., the $ddTTP$ terminated products) and simultaneously analyzing the pooled products by mass spectrometry. Again, the first extension products of the known nucleic acid primer sequence are assigned, e.g., via a computer program. Mass/sequence assignments are possible even in the worst case in which the nucleic acid primer is extended/terminated by the same nucleotide, e.g., ddT , in all three clones. The following configurations thus obtained can be well differentiated by their different mass modifications: $UP^0 ddT^0$, $UP^0 ddT^1$, $UP^0 ddT^2$.

In yet another embodiment of this invention, DNA sequencing by multiplex mass spectrometry can be achieved by cloning the DNA fragments to be sequenced in "plex-vectors" containing vector specific "tag sequences" as described (Köster et al., "Oligonucleotide Synthesis and Multiplex DNA Sequencing Using Chemiluminescent Detection," *Nucleic Acids Re. Symposium Ser. No. 24*, 318-321 (1991)); then pooling clones from different plex-vectors for DNA preparation and the four separate Sanger sequencing reactions using standard $dNTP^0/ddNTP^0$ and nucleic acid primer UP^0

purifying the four multiplex fragment families via linking to a solid support through the linking group, L, at the 5'-end of UP⁰; washing out all by-products, and cleaving the purified multiplex DNA fragments off the support or using the L-L' bound nested Sanger fragments as such for mass spectrometric analysis as described above; performing de-multiplexing by one-by-one hybridization of specific "tag probes"; and subsequently analyzing by mass spectrometry. As a reference point, the four base-specifically terminated multiplex DNA fragment families are run by the mass spectrometer and all ddT⁰, ddA ddC and ddG⁰ terminated molecular ion peaks are respectively detected and memorized. Assignment of, for example, ddT⁰ terminated DNA fragments to a specific fragment family is accomplished by another mass spectrometric analysis after hybridization of the specific tag probe (TP) to the corresponding tag sequence contained in the sequence of this specific fragment family.

Only those molecular ion peaks which are capable of hybridizing to the specific tag probe are shifted to a higher molecular mass by the same known mass increment (e.g. of the tag probe). These shifted ion peaks, by virtue of all hybridizing to a specific tag probe, belong to the same fragment family. For a given fragment family, this is repeated for the remaining chain terminated fragment families with the same tag probe to assign the complete DNA sequence. This process is repeated *i*-1 times corresponding to *i* clones multiplexed (the *i*-th clone is identified by default).

The differentiation of the tag probes for the different multiplexed clones can be obtained just by the DNA sequence and its ability to Watson-Crick base pair to the tag sequence. It is well known in the art how to calculate stringency conditions to provide for specific hybridization of a given tag probe with a given tag sequence (see, for example, *Molecular Cloning: A laboratory manual* 2ed, ed. by Sambrook, Fritsch and Maniatis (Cold Spring Harbor Laboratory Press: NY, 1989, Chapter 11). Furthermore, differentiation can be obtained by designing the tag sequence for each plex-vector to have a sufficient mass difference so as to be unique just by changing the length or base composition or by mass-modifications. In order to keep the duplex between the tag sequence and the tag probe intact during mass spectrometric analysis, it is another embodiment of the invention to provide for a covalent attachment mediated by, for example, photoreactive groups such as psoralen and ellipticine and by other methods known to those skilled in the art (see, for example, Hélène et al.,

Nature 344, 358 (1990) and Thuong et al. "Oligonucleotides Attached to Intercalators, Photoreactive and Cleavage Agents" in F. Eckstein, Oligonucleotides and Analogues A Practical Approach, IRL Press, Oxford 1991, 283-306).

The DNA sequence is unraveled again by searching for the lowest molecular weight molecular ion peak corresponding to the known UP⁰-tag sequence/tag probe molecular weight plus the first extension product, e.g., ddT⁰, then the second, the third, etc.

In a combination of the latter approach with the previously described multiplexing processes, a further increase in multiplexing can be achieved by using, in addition to the tag probe/tag sequence interaction, mass-modified nucleic acid primers and/or mass-modified deoxynucleoside, dNTP⁰⁻ⁱ, and/or dideoxynucleoside triphosphates, ddNTP⁰⁻ⁱ. Those skilled in the art will realize that the tag sequence/tag probe multiplexing approach is not limited to Sanger DNA sequencing generating nested DNA fragments with DNA polymerases. The DNA sequence can also be determined by transcribing the unknown DNA sequence from appropriate promoter-containing vectors (see above) with various RNA polymerases and mixtures of NTP⁰⁻¹ 3' dNTP⁰⁻¹, thus generating nested RNA fragments.

In yet another embodiment of this invention, the mass-modifying functionality can be introduced by a two or multiple step process. In this case, the nucleic acid primer, the chain-elongating or terminating nucleoside triphosphates and/or the tag probes are, in a first step, modified by a precursor functionality such as azido, -N₃, or modified with a functional group in which the R in XR is H thus providing temporary functions, e.g., but not limited to -OH, -NH₂, -NHR, -SH, -NCS, -OCO(CH₂)_rCOOH (r = 1-20), -NHCO(CH₂)_rCOOH (r = 1-20), -OSO₂OH, -OCO(CH₂)_r' (r = 1-20), -OP(O-Alkyl)N(Alkyl)₂. These less bulky functionalities result in better substrate properties for the enzymatic DNA or RNA synthesis reactions of the DNA sequencing process. The appropriate mass-modifying functionality is then introduced after the generation of the nested base-specifically terminated DNA or RNA fragments prior to mass spectrometry. Several examples of compounds which can serve as mass-modifying functionalities are depicted herein without limiting the scope of this invention.

1.4.1.6 Kits for Sequencing Nucleic Acid by Mass Spectrometry

Another aspect of this invention concerns kits for sequencing nucleic acids by mass spectrometry which include combinations of the above-described sequencing

reactants. For instance, in one embodiment, the kit comprises reactants for multiplex mass spectrometric sequencing of several different species of nucleic acid. The kit can include a solid support having a linking functionality (L 1) for immobilization of the base- specifically terminated products; at least one nucleic acid primer having a linking group (L) for reversibly and temporarily linking the primer and solid support through, for example, a photocleavable bond; a set of chain-elongating nucleotides (e. g., dATP, dCTP, dGTP and dTTP, or ATP, CTP, GTP and UTP); a set of chain-terminating nucleotides (such as 2',3'-dideoxynucleotides for DNA synthesis or 3' deoxynucleotides for RNA synthesis); and an appropriate polymerase for synthesizing complementary nucleotides. Primers and/or terminating nucleotides can be mass-modified so that the base-specifically terminated fragments generated from one of the species of nucleic acids to be sequenced can be distinguished by mass spectrometry from all of the others. Alternative to the use of mass-modified synthesis reactants, a set of tag probes (as described above) can be included in the kit. The kit can also include appropriate buffers as well as instructions for performing multiplex mass spectrometry to concurrently sequence multiple species of nucleic acids.

In another embodiment, a nucleic acid sequencing kit can comprise a solid support as described above, a primer for initiating synthesis of complementary nucleic acid fragments, a set of chain-elongating nucleotides and an appropriate polymerase. The mass-modified chain-terminating nucleotides are selected so that the addition of one of the chain terminators to a growing complementary nucleic acid can be distinguished by mass spectrometry.

1.4.2 A Method And System For Determining The Sequence Of Genomes

1.4.2.1 A Process For Directly Amplifying And Base Specifically Terminating A Nucleic Acid Molecule For Sequencing

In general., the invention features a process for directly amplifying and base specifically terminating a nucleic acid molecule. According to the process of the invention, a combined amplification and termination reaction is performed on a nucleic acid template using: i) a complete set of chain-elongating nucleotides; ii) at least one chain-terminating nucleotide; and (iii) a first DNA polymerase, which has a relatively low affinity towards the chain terminating nucleotide; and (iv) a second DNA polymerase, which has a relatively high affinity towards the chain terminating nucleotide, so that polymerization by the enzyme with relatively low affinity for the chain terminating nucleotide leads to amplification of the template, whereas the enzyme with relatively high affinity for the chain terminating nucleotide terminates the polymerization and yields sequencing products.

The combined amplification and sequencing can be based on any amplification procedure that employs an enzyme with polynucleotide synthetic ability (e.g. polymerase). One preferred process, based on the polymerase chain reaction (PCR), is comprised of the following three thermal steps: 1) denaturing a double stranded (ds) DNA molecule at an appropriate temperature and for an appropriate period of time to obtain the two single stranded (ss) DNA molecules (the template: sense and antisense strand); 2) contacting the template with at least one primer that hybridizes to at least one ss DNA template at an appropriate temperature and for an appropriate period of time to obtain a primer containing ss DNA template; 3) contacting the primer containing template at an appropriate temperature and for an appropriate period of time with: (i) a complete set of chain elongating nucleotides, (ii) at least one chain terminating nucleotide, (iii) a first DNA polymerase, which has a relatively low affinity towards the chain terminating nucleotide; and (iv) a second DNA polymerase, which has a relatively high affinity towards the chain terminating nucleotide.

Steps 1)- 3) can be sequentially performed for an appropriate number of times (cycles) to obtain the desired amount of amplified sequencing ladders. The quantity of the base specifically terminated fragment desired dictates how many cycles are performed. Although an increased number of cycles results in an increased level of

amplification, it may also detract from the sensitivity of a subsequent detection. It is therefore generally undesirable to perform more than about 50 cycles, and is more preferable to perform less than about 40 cycles (e.g. about 20-30 cycles). In a preferred embodiment, the first denaturation step is performed at a temperature in the range of about 85°C to about 100°C (most preferably about 92°C to about 96°C) for about 20 seconds (s) to about 2 minutes (most preferably about 30s- 1 minute). The second hybridization step is preferably performed at a temperature, which is in the range of about 40°C to about 80°C (most preferably about 45°C to about 72°C) for about 20s to about 2 minutes (most preferably about 30s-1 minute). The third, primer extension step is preferably performed at about 65°C to about 80°C (most preferably about 70°C to about 74°C) for about 30 s to about 3 minutes (most preferably about 1 to about 2 minutes).

In order to obtain sequence information on both the sense and antisense strands of a DNA molecule simultaneously, each of the single stranded sense and antisense templates generated from the denaturing step can be contacted with appropriate primers in step 2), so that amplified and chain terminated nucleic acid molecules generated in step 3), are complementary to both strands.

Another preferred process for simultaneously amplifying and chain terminating a nucleic acid sequence is based on strand displacement amplification (SDA) (G. Terrance Walker et al., *Nucleic Acids Res.* 22, 2670-77 (1994); European Patent Publication Number 0 684 315 entitled Strand Displacement Amplification Using Thermophilic Enzymes). In essence, this process involves the following three steps, which altogether comprise a cycle: 1) denaturing a double stranded (ds) DNA molecule containing the sequence to be amplified at an appropriate temperature and for an appropriate period of time to obtain the two single stranded (ss) DNA molecules (the template: sense and antisense strand); 2) contacting the template with at least one primer (P), that contains a recognition/cleavage site for a restriction endonuclease (RE) and that hybridizes to at least one ss DNA template at an appropriate temperature and for an appropriate period of time to obtain a primer containing ss DNA template; 3) contacting the primer containing template at an appropriate temperature and for an appropriate period of time with: (i) a complete set of chain elongating nucleotides; (ii) at least one chain terminating nucleotide, (iii) a first DNA polymerase, which has a relatively low affinity towards the chain

terminating nucleotide; (iv) a second DNA polymerase, which has a relatively high affinity towards the chain terminating nucleotide; and (v) an RE that nicks the primer recognition/cleavage site.

Steps 1) - 3) can be sequentially performed for an appropriate number of times (cycles) to obtain the desired amount of amplified sequencing ladders. As with the PCR based process, the quantity of the base specifically terminated fragment desired dictates how many cycles are performed. Preferably, less than 50 cycles, more preferably less than about 40 cycles and most preferably about 20 to 30 cycles are performed.

The amplified sequencing ladders obtained as described above, can be separated and detected and/or quantitated using well established methods, such as polyacrylamide gel electrophoresis (PAGE), or capillary zone electrophoresis (CZE) (Jorgenson et al., *J. Chromatography* 352, 337 (1986); Gesteland et al., *Nucleic Acids Res.* L8, 1415-1419 (1990)); or direct blotting electrophoresis (DBE) (Beck and Pohl, *EMBO J*, vol. 3: Pp. 2905-2909 (1984)) in conjunction with, for example, colorimetry, fluorimetry, chemiluminescence and radioactivity.

Dye-terminator chemistry can be employed in the combined amplification and sequencing reaction to enable the simultaneous generation of forward and reverse sequence ladders, which can be separated based on the streptavidin-biotin system when one biotinylated primer is provided.

Depicted herein is a scheme for the combined amplification and sequencing using two polymerases and dye-labeled chain terminating nucleotide (ddNTP) for detection and two reverse oriented primers. A means of separation for the simultaneously generated forward and reverse sequence ladders is shown. Step A represents the exponential amplification of a target sequence by the polymerase with a low affinity for ddNTPs. One of the sequence specific oligonucleotide primers is biotinylated. Step B represents the generation of a sequence ladder either from the original template or the simultaneously generated amplification product carried out by the polymerase with a high affinity for ddNTPs. After completion of the reaction, the products are incubated with a streptavidin coated solid support (Step C). Biotinylated forward sequencing products and reverse products hybridized to the forward template are immobilized. In order to obtain readable sequence information, the forward and reverse sequence ladders are separated in Step D. The immobilized strands are

washed and separated by denaturation with ammonium hydroxide at room temperature. The non-biotinylated reverse sequencing products are removed from the beads with ammonium hydroxide supernatant during this procedure. The biotinylated forward sequencing products remain immobilized to the beads and are re-solubilized with ammonium hydroxide at 60°C. After ethanol precipitation, both sequencing species can be resuspended in loading dye and run on an automated sequencer, for example.

When mass spectrometry is used in conjunction with the direct amplification and chain termination processes, the sequencing ladders can be directly detected without first being separated using several mass spectrometer formats.

Amenable formats for use in the invention include ionization techniques such as matrix-assisted laser desorption (MALDI), continuous or pulsed electrospray (ESI) and related methods (e.g. Ionspray or Thermospray), and massive cluster impact (MSI); these ion sources can be matched with a detection format, such as linear or reflectron time-of-flight (TOF), single or multiple quadrupole, single or multiple magnetic sector, Fourier Transform ion cyclotron resonance (FTICR), ion trap, or combinations of these to give a hybrid detector (e.g. ion trap-TOF). For ionization, numerous matrix/wavelength combinations (MALDI) or solvent combinations (ESI) can be employed.

The above-described process can be performed using virtually any nucleic acid molecule as the source of the DNA template. For example, the nucleic acid molecule can be: a) single stranded or double stranded; b) linear or covalently closed circular in supercoiled or relaxed form; or c) RNA if combined with reverse transcription to generate a cDNA. For example, reverse transcription can be performed using a suitable reverse transcriptase (e.g. Moloney murine leukemia virus reverse transcriptase) using standard techniques (e.g. Kawasaki (1990) in PCR Protocols: A Guide to Methods and Applications, Innis et al., eds., Academic Press, Berkeley, CA pp21- 27).

Sources of nucleic acid templates can include: a) plasmids (naturally occurring or recombinant); b) RNA- or DNA- viruses and bacteriophages (naturally occurring or recombinant); c) chromosomal or episomal replicating DNA (e. g. from tissue, a blood sample, or a biopsy); d) a nucleic acid fragment (e.g. derived by exonuclease, unspecific endonuclease or restriction endonuclease digestion or by physical

disruption (e.g. sonication or nebulization)); and e) RNA or RNA transcripts like mRNAs.

The nucleic acid to be amplified and sequenced can be obtained from virtually any biological sample. As used herein, the term "biological sample" refers to any material obtained from any living source (e.g. human, animal, plant, bacteria, fungi, protist, virus). Examples of appropriate biological samples for use in the instant invention include: solid materials (e.g. tissue, cell pellets, biopsies) and biological fluids (e.g. urine, blood, saliva, amniotic fluid, mouth wash, spinal fluid). The nucleic acid to be amplified and sequenced can be provided by unpurified whole cells, bacteria or virus.

Alternatively, the nucleic acid can first be purified from a sample using standard techniques, such as: a) cesium chloride gradient centrifugation; b) alkaline lysis with or without RNase treatment; c) ion exchange chromatography; d) phenol/chloroform extraction; e) isolation by hybridization to bound oligonucleotides; f) gel electrophoresis and elution; alcohol precipitation and h) combinations of the above.

As used herein, the phrases "chain-elongating nucleotides" and "chain-terminating nucleotides" are used in accordance with their art recognized meaning. For example, for DNA, chain-elongating nucleotides include 2'-deoxyribonucleotides (e.g. dATP, dCTP, dGTP and dTTP) and chain-terminating nucleotides include 2', 3'-dideoxyribonucleotides, (e.g. ddATP, ddCTP, ddGTP, ddTTP). For RNA, chain-elongating nucleotides include ribonucleotides (e.g., ATP, CTP, GTP and UTP) and chain-terminating nucleotides include 3'-deoxyribonucleotides (e.g. 3'dA, 3'dC, 3'dG and 3'dU). A complete set of chain elongating nucleotides refers to dATP, dCTP, dGTP and dTTP. The term "nucleotide" is also well known in the art. For the purposes of this invention, nucleotides include nucleoside mono-, di-, and triphosphates. Nucleotides also include modified nucleotides, such as phosphorothioate nucleotides and deazapurine nucleotides. A complete set of chain-elongating nucleotides refers to four different nucleotides that can hybridize to each of the four different bases comprising the DNA template.

If the amplified sequencing ladders are to be detected by mass spectrometric analysis, it may be useful to "condition" nucleic acid molecules, for example to decrease the laser energy required for volatilization and/or to minimize fragmentation.

Conditioning is preferably performed while the sequencing ladders are immobilized. An example of conditioning is modification of the phosphodiester backbone of the nucleic acid molecule (e.g. cation "change"), which can be useful for eliminating peak broadening due to a heterogeneity in the cations bound per nucleotide unit.

Contacting a nucleic acid molecule, which contains an -thio-nucleoside-triphosphate during polymerization with an alkylating agent such as alkyl iodide, iodoacetamide, - iodoethanol, or 2,3-epoxy-1-propanol, the monothio phosphodiester bonds of a nucleic acid molecule can be transformed into a phosphotriester bond. Further conditioning involves incorporating nucleotides which reduce sensitivity for depurination (fragmentation during MS), e.g. a purine analog such as N7- or N9-deazapurine nucleotides, and partial RNA containing oligodeoxynucleotide to be able to remove the unmodified primer from the amplified and modified sequencing ladders by RNase or alkaline treatment. In DNA sequencing using fluorescent detection and gel electrophoretic separation, the N7 deazapurine nucleotides reduce the formation of secondary structure resulting in band compression from which no sequencing information can be generated.

1.4.2.2 The Use of Two Polymerase Enzymes Each Having Different Affinities for the Chain Terminating Nucleotides

Critical to the novel process of the invention is the use of appropriate amounts of two different polymerase enzymes, each having a different affinity for the particular chain terminating nucleotide, so that polymerization by the enzyme with relatively low affinity for the chain terminating nucleotide leads to amplification whereas the enzyme with relatively high affinity for the chain terminating nucleotide terminates the polymerization and yields sequencing products. Preferably about 0.5 to about 3 units of polymerase is used in the combined amplification and chain termination reaction. Most preferably about 1 to 2 units is used. Particularly preferred polymerases for use in conjunction with PCR or other thermal amplification process are thermostable polymerases, such as Taq DNA polymerase (Boehringer Mannheim), AmpliTaq FS DNA polymerase (Perkin-Elmer), Deep Vent (exo-), Vent, Vent (exo-) and Deep Vent DNA polymerases (New England Biolabs), Thermo Sequenase (Amersham) or exo(-) Pseudococcus furiosus (Pfu) DNA polymerase (Stratagene, Heidelberg Germany). AmpliTaq, Ultman, 9 degree Nm, Tth, Hot Tub,

and *Pyrococcus furiosus*. In addition, preferably the polymerase does not have 5'-3' exonuclease activity.

The process of the invention can be carried out using AmpliTaq FS DNA polymerase (Perkin-Elmer), which has a relatively high affinity and Taq DNA polymerase, which has a relatively low affinity for chain terminating nucleotides. Other appropriate polymerase pairs for use in the instant invention can be determined by one of skill in the art. (See e.g. S. Tabor and C.C. Richardson (1995) Proc. Nat. Acad. Sci. (USA), vol. 92: Pp. 6339-6343.) in addition to polymerases, which have a relatively high and a relatively low affinity to the chain terminating nucleotide, a third polymerase, which has proofreading capacity (e.g. *Pyrococcus woesei* (Pwo)) DNA polymerase may also be added to the amplification mixture to enhance the fidelity of amplification.

Oligonucleotide primers, for use in the invention, can be designed based on knowledge of the 5' and/or 3' regions of the nucleotide sequence to be amplified and sequenced, e.g., insert flanking regions of cloning and sequencing vectors (such as M13, pUC, phagemid, costaid). Optionally, at least one primer used in the chain extension and termination reaction can be linked to a solid support to facilitate purification of amplified product from primers and other reactants, thereby increasing yield or to separate the Sanger ladders from the sense and antisense template strand where simultaneous amplification-sequencing of both a sense and antisense strand of the template DNA has been performed.

Examples of appropriate solid supports include beads (silica gel, controlled pore glass, magnetic beads, Sephadex/Sepharose beads, cellulose beads, etc.), capillaries, flat supports such as glass fiber filters, glass surfaces, metal surfaces (steel, gold, silver, aluminum, and copper), plastic materials or membranes (polyethylene, polypropylene, polyamide, polyvinylidenedifluoride) or beads in pits of flat surfaces such as wafers (e.g. silicon wafers), with or without filter plates.

1.4.2.3 Immobilization Based on Hybridization

Immobilization can be accomplished, for example, based on hybridization between a capture nucleic acid sequence, which has already been immobilized to the support and a complementary nucleic acid sequence, which is also contained within the nucleic acid molecule containing the nucleic acid sequence to be detected. So that hybridization between the complementary nucleic acid molecules is not hindered by

the support, the capture nucleic acid can include a spacer region of at least about five nucleotides in length between the solid support and the capture nucleic acid sequence. The duplex formed will be cleaved under the influence of the laser pulse and desorption can be initiated. The solid support-bound base sequence can be presented through natural oligoribo- or oligodeoxyribo- nucleotide as well as analogs (e.g. thio-modified phosphodiester or phosphotriester backbone) or employing oligonucleotide mimetics such as PNA analogs (see e.g. Nielsen et al., Science, 254, 1497 (1991)) which render the base sequence less susceptible to enzymatic degradation and hence increases overall stability of the solid support-bound capture base sequence.

1.4.2.4 Linkage

Alternatively, a target detection site can be directly linked to a solid support via a reversible or irreversible bond between an appropriate functionality (L') on the target nucleic acid molecule and an appropriate functionality (L) on the capture molecule. A reversible linkage can be such that it is cleaved under the conditions of mass spectrometry (i.e., a photocleavable bond such as a trityl ether bond or a charge transfer complex or a labile bond being formed between relatively stable organic radicals). Furthermore, the linkage can be formed with L' being a quaternary ammonium group, in which case, preferably, the surface of the solid support carries negative charges which repel the negatively charged nucleic acid backbone and thus facilitate the desorption required for analysis by a mass spectrometer. Desorption can occur either by the heat created by the laser pulse and/or, depending on L', by specific absorption of laser energy which is in resonance with the L' chromophore.

By way of example, the L-L' chemistry can be of a type of disulfide bond (chemically cleavable, for example, by mercaptoethanol or dithioerythrol), a biotin/streptavidin system, a heterobifunctional derivative of a trityl ether group (Köster et al., "A Versatile Acid-Labile Linker for Modification of Synthetic Biomolecules," Tetrahedron Letters 31, 7095 (1990)) which can be cleaved under mildly acidic conditions as well as under conditions of mass spectrometry, a levulinyl group cleavable under almost neutral conditions with a hydrazinium/acetate buffer, an arginine-arginine or lysine-lysine bond cleavable by an endopeptidase enzyme like trypsin or a pyrophosphate bond cleavable by a pyrophosphatase or a ribonucleotide in between a deoxynucleotide sequence cleavable by an RNase or alkali.

The functionalities, L and L', can also form a charge transfer complex and thereby form the temporary L-L' linkage. Since in many cases the "charge-transfer band" can be determined by UV/vis spectrometry (see e.g. Organic Charge Transfer Complexes by R. Foster, Academic Press, 1969), the laser energy can be tuned to the corresponding energy of the charge-transfer wavelength and, thus, a specific desorption off the solid support can be initiated. Those skilled in the art will recognize that several combinations can serve this purpose and that the donor functionality can be either on the solid support or coupled to the nucleic acid molecule to be detected or vice versa.

In yet another approach, a reversible L-L' linkage can be generated by homolytically forming relatively stable radicals. Under the influence of the laser pulse, desorption (as discussed above) as well as ionization will take place at the radical position. Those skilled in the art will recognize that other organic radicals can be selected and that, in relation to the dissociation energies needed to homolytically cleave the bond between them, a corresponding laser wavelength can be selected (see e.g. Reactive Molecules by C. Wentrup, John Wiley & Sons, 1984). An anchoring function L' can also be incorporated into a target capturing sequence by using appropriate primers during an amplification procedure, such as PCR, LCR or transcription amplification.

For certain applications, it may be useful to simultaneously amplify and chain terminate more than one (mutated) loci on a particular captured nucleic acid fragment (on one spot of an array) or it may be useful to perform parallel processing by using oligonucleotide or oligonucleotide mimetic arrays on various solid supports.

"Multiplexing" can be achieved either by the sequence itself (composition or length) or by the introduction of mass-modifying functionalities into the primer oligonucleotide. Such multiplexing is particularly useful in conjunction with mass spectrometric DNA sequencing or mobility modified gel based fluorescence sequencing.

1.4.2.5 Mass or Mobility Modification

Without limiting the scope of the invention, the mass or mobility modification can be introduced by using oligo/polyethylene glycol derivatives. The oligo/polyethylene glycols can also be monoalkylated by a lower alkyl such as methyl, ethyl, propyl, isopropyl, t-butyl and the like. Other chemistries can be used in

the mass-modified compounds, as for example, those described recently in *Oligonucleotides and Analogues- A Practical Approach*, F. Eckstein, editor IRL Press, Oxford, 1991.

In yet another embodiment, various mass or mobility modifying functionalities, other than oligo/polyethylene glycols, can be selected and attached via appropriate linking chemistries. A simple modification can be achieved by using different alkyl, aryl or aralkyl moieties such as methyl, ethyl, propyl, isopropyl, t-butyl, hexyl, phenyl, substituted phenyl or benzyl. Yet another modification can be obtained by attaching homo- or heteropeptides to the nucleic acid molecule (e.g., primer) or nucleoside triphosphates. Simple oligoamides also can be used. Numerous other possibilities, in addition to those mentioned above, can be performed by one skilled in the art.

Different mass or mobility modified primers allow for multiplex sequencing via simultaneous detection of primer-modified Sanger sequencing ladders.

Mass or mobility modifications can be incorporated during the amplification process through nucleoside triphosphates or modified primers.

1.4.2.6 Kits for Amplified Base Specifically Terminated Fragments

Another aspect of this invention concerns kits for directly generating from a nucleic acid template, amplified base specifically terminated fragments. Such kits include combinations of the above-described reactants. For instance, in one embodiment, the kit can comprise: i) a set of chain-elongating nucleotides; ii) a set of chain-terminating nucleotides; and (iii) a first DNA polymerase, which has a relatively low affinity towards the chain terminating nucleotide; and (iv) a second DNA polymerase, which has a relatively high affinity towards the chain terminating nucleotide. The kit can also include appropriate solid supports for capture/purification and buffers as well as instructions for use.

For use with certain detection means, such as polyacrylamide gel electrophoresis (PAGE), detectable labels must be used in either the primer (typically at the 5'-end) or in one of the chain extending nucleotides, or chain terminating nucleotides.

Using radioisotopes such as ^{32}P , ^{33}P , or ^{31}S is still the most frequently used technique.

After PAGE, the gels are exposed to X-ray films and silver grain exposure is analyzed.

1.4.3 Hybridization

Oligonucleotide arrays can be used in a wide variety of applications, including hybridization studies. In a hybridization study, the array can be exposed to a receptor (R) of interest. The receptor can be labelled with an appropriate label (*), such as fluorescein. The locations on the substrate where the receptor has bound are determined and, through knowledge of the sequence of the oligonucleotide probe at that location one can then determine, if the receptor is an oligonucleotide, the sequence of the receptor.

Sequencing by hybridization (SBH) is most efficiently practiced by attaching many probes to a surface to form an array in which the identity of the probe at each site is known. A labeled target DNA or RNA is then hybridized to the array, and the hybridization pattern is examined to determine the identity of all complementary probes in the array. Contrary to the teachings of the prior art, which teaches that mismatched probe/target complexes are not of interest, the present invention provides an analytical method in which the hybridization signal of mismatched probe/target complexes identifies or confirms the identity of the perfectly matched probe/target complexes on the array.

Arrays of oligonucleotides are efficiently generated for the hybridization studies using light-directed synthesis techniques.

1.4.3.1 Light Directed Synthesis

As discussed below, an array of all tetranucleotides was produced in sixteen cycles, which required only 4 hours to complete. Because combinatorial strategies are used, the number of different compounds on the array increases exponentially during synthesis, while the number of chemical coupling cycles increases only linearly. For example, expanding the synthesis to the complete set of 4^8 (65,536) octanucleotides adds only 4 hours (or less) to the synthesis due to the 16 additional cycles required. Furthermore, combinatorial synthesis strategies can be implemented to generate arrays of any desired probe composition. For example, because the entire set of dodecamers (4^{12}) can be produced in 48 photolysis and coupling cycles or less (b^n compounds requires no more than $b \times n$ cycles), any subset of the dodecamers (including any subset of shorter oligonucleotides) can be constructed in 48 or fewer chemical coupling steps. The number of compounds in an array is limited only by the density of synthesis sites and the overall array size. The present invention has been

practiced with arrays with probes synthesized in square sites 25 microns on a side. At this resolution, the entire set of 65,536 octanucleotides can be placed in an array measuring only 0.64 cm². The set of 1,048,576 dodecanucleotides requires only a 2.56 cm² array at this individual probe site size.

The success of genome sequencing projects depends on efficient DNA sequencing technologies. Current methods are highly reliant on complex procedures and require substantial manual effort. SBH offers the potential for automating many of the manual efforts in current practice. Light-directed synthesis offers an efficient means for large scale production of miniaturized arrays not only for SBH but for many other applications as well.

Although oligonucleotide arrays can be used for primary sequencing applications, many diagnostic methods involve the analysis of only a few nucleotide positions in a target nucleic acid sequence. Because single base changes cause multiple changes in the hybridization pattern of the target on a probe array, the oligonucleotide arrays and methods of the present invention enable one to check the accuracy of previously elucidated DNA sequences, or to scan for changes or mutations in certain specific sequences within a target nucleic acid. The latter as is important, for example, for genetic disease, quality control, and forensic analysis. With an octanucleotide probe set, a single base change in a target nucleic acid can be detected by the loss of eight perfect hybrids, and the generation of eight new perfect hybrids. The single base change can also be detected through altered mismatch probe/target complex formation on the array. Perhaps even more surprisingly, such single base changes in a complex nucleic acid dramatically alter the overall hybridization pattern of the target to the array. According to the present invention such changes in the overall hybridization pattern are used to actually simplify the analysis.

The high information content of light-directed oligonucleotide arrays greatly benefits genetic diagnostic testing. Sequence comparisons of hundreds to thousands of different mutations can be assayed simultaneously instead of in a one-at-a-time format.

1.4.3.2 Arrays Constructed to Contain Genetic Markers

Arrays can also be constructed to contain genetic markers for the rapid identification of a wide variety of pathogenic organisms, and to study the sequence

specificity of RNA/RNA, RNA/DNA, protein/RNA or protein/DNA, interactions. One can use non Watson- Crick oligonucleotides and novel synthetic nucleoside analogs for antisense, triple helix, or other applications. Suitably protected RNA monomers can be employed for RNA synthesis, and a wide variety of synthetic and non-naturally occurring nucleic acid analogues can be used, depending upon the motivations of the practitioner. See, e.g., PCT patent Publication Nos. 91/19813, 92/05285, and 92/14843, incorporated herein by reference. In addition, the oligonucleotide arrays can be used to deduce thermodynamic and kinetic rules governing the formation and stability of oligonucleotide complexes.

1.4.3.2.1 Hybridization of Targets to Surface Oligonucleotides

The support bound octanucleotide probes discussed above were hybridized to a target of 5'GCGTAGGC-fluorescein in the hybridization chamber by incubation for 15 minutes at 15°C.

The array surface was then interrogated with an epifluorescence microscope (488 nm argon ion excitation). The fluorescence intensity pattern matches the 800 X 1280 μm stripe used to direct the synthesis of the probe. Furthermore, the signal intensities are high (four times over the background of the glass substrate), demonstrating specific binding of the target to the probe.

The behavior of the target-probe complex was investigated by increasing the temperature of the hybridization solution. After a minute equilibration at each temperature, the substrate was scanned for signal. The duplex melted in the temperature range expected for the sequence under study ($T_m \sim 28^\circ\text{C}$ obtained from the rule $T_m = [2^\circ(A+T) + 4^\circ(G+C)]$). The probes in the array were stable to temperature denaturation of the target-probe complex as demonstrated by rehybridization of target DNA.

1.4.3.2.2 Sequence Specificity of Target Hybridization

To demonstrate the sequence specificity of target hybridization, two different probes were synthesized in 800 x 1280 μm stripes. The probe S-3'-CGCATCCG was synthesized in stripes 1, 3 and 5. The probe S-3'-CGCTTCCG was synthesized in stripes 2, 4 and 6. The results of hybridizing a 5'-GCGTAGGC-fluorescein target to the substrate at 15°C are depicted herein. Although the probes differ by only one internal base, the target hybridizes specifically to its

complementary sequence (~500 counts above background in stripes 1, 3 and 5) with little or no detectable signal in positions 2, 4 and 6 (~10 counts).

1.4.3.2.3 Combinatorial Synthesis of, and Hybridization of a Nucleic Acid Target to, a Probe Matrix

In a light-directed synthesis, the location and composition of products depends on the pattern of illumination and the order of chemical coupling reagents (see Fodor et al., *Science* (1991) 251:767-773, for a complete description). Consider the synthesis of 256 tetranucleotides. Mask 1 activates one fourth of the substrate surface for coupling with the first of four nucleosides in the first round of synthesis. In cycle 2, mask 2 activates a different quarter of the substrate for coupling with the second nucleoside. The process is continued to build four regions of mononucleotides. The masks of round 2 are perpendicular to those of round 1, and each cycle of round 2 generates four new dinucleotides. The process continues through round 2 to form sixteen dinucleotides. The masks of round 3 further subdivide the synthesis regions so that each coupling cycle generates 16 trimers. The subdivision of the substrate is continued through round 4 to form the tetranucleotides. The synthesis of this probe matrix can be compactly represented in polynomial notation as $(A+C+G+T)^4$. Expansion of this polynomial yields the 256 tetranucleotides.

The application of an array of 256 probes synthesized by light-directed combinatorial synthesis to generate a probe matrix is illustrated herein. The polynomial for this synthesis is given by: $3'-CG(A+G+C+T)^4CG$. All possible tetranucleotides were synthesized flanked by CG at the 3'- and 5'-ends. Hybridization of target 5'-GCGGCGGC-fluorescein to this array at 15°C correctly yielded the S-3'-CGCCGCCG complementary probe as the most intense position (2,698 counts). Significant intensity was also observed for the following mismatches: S-3'-CGCAGCCG (554 counts), S-3'-CGCCGACG (317 counts), S-3'-CGCCGTCTG (272 counts), S-3'-CGACGCCG (242 counts), S-3'-CGTCGCCG (203 counts), S-3'-CGCCCCCG (180 counts), S-3'-CGCTGCCG (163 counts), S-3'-CGCCACCG (125 counts), and S-3'-CGCCTCCG (78 counts).

1.4.3.3 Mismatch Analysis

1.4.3.3.1 Arrays Used to Determine the Gene Sequence of Oligos of Length "n" Using Array of Probes of Shorter Length "k"

The arrays discussed herein can be utilized in the present method to determine the nucleic acid sequence of an oligonucleotide of length n using an array of probes of shorter length k . In a simple example, the target has a sequence 5'-XXYXY-3', where X and Y are complementary nucleic acids such as A and T or C and G. For discussion purposes, the example is simplified by using only two bases and very short sequences, but the technique can easily be extended to larger nucleic acids with, for example, all 4 RNA or DNA bases.

The sequence of the target is, generally, not known ab initio. One can determine the sequence of the target using the present method with an array of shorter probes. In this example, an array of all possible X and Y 4-mers is synthesized and then used to determine the sequence of a 5-mer target.

Initially, a "core" probe is identified. The core probe is exactly complementary to a sequence in the target using the mismatch analysis method of the present invention. The core probe is identified using one or both of the following criteria:

1. The core probe exhibits stronger binding affinity to the target than other probes, typically the strongest binding affinity of any probe in the array (that has not been identified as a core probe in a previous cycle of analysis).
2. Probes that are mismatched with the target, as compared to the core probe sequence, exhibit a characteristic pattern, discussed in greater detail below, in which probes that mismatch at the 3'- and 5'-end of the probe bind more strongly to the target than probes that mismatch at interior positions.

In this particular example, selection criteria #1 identifies a core 4-mer probe with the strongest binding affinity to the target that has the sequence 3'-YYXY. The probe 3'-YYXY (corresponding to the 5'-XXYX position of the target) is, therefore, chosen as the "core" probe.

Selection criteria #2 is utilized as a "check" to ensure the core probe is exactly complementary to the target nucleic acid.

The second selection criteria evaluates hybridization data (such as the fluorescence intensity of a labeled target hybridized to an array of probes on a

substrate, although other techniques are well known to those of skill in the art) of probes that have single base mismatches as compared to the core probe. In this particular case, the core probe has been selected as S-3'-YYXY. The single base mismatched probes of this core probe are: S-3'-XYXY, S-3'-YXXY, S-3'-YYYY, and S-3'-YYXX. The binding affinity characteristics of these single base mismatches are utilized to ensure that a "correct" core has been selected, or to select the core probe from among a set of probes exhibiting similar binding affinities.

1.4.3.3.2 Binding Affinity vs. Mismatch Position

An illustrative, hypothetical plot of expected binding affinity versus mismatch position is provided herein. The binding affinity values (typically fluorescence intensity of labeled target hybridized to probe, although many other factors relating to affinity may be utilized) are all normalized to the binding affinity of S-3'-YYXY to the target, which is plotted as a value of 1. Because only two nucleotides are involved in this example, the value plotted for a probe mismatched at position 1 (the nucleotide at the 3'-end of the probe) is the normalized binding affinity of S-3'-XYXY. The value plotted for mismatch at position 2 is the normalized affinity of S-3'-YXXY. The value plotted for mismatch at position 3 is the normalized affinity of S-3'-YYYY, and the value plotted for mismatch position 4 is the normalized affinity of S-3'-YYXX. As noted above, "affinity" may be measured in a number of ways including, for example, the number of photon counts from fluorescence markers on the target.

The affinity of all three mismatches is lower than the core in this illustration. Moreover, the affinity plot shows that a mismatch at the 3'-end of the probe has less impact than a mismatch at the 5'-end of the probe in this particular case, although this may not always be the case. Further, mismatches at the end of the probe result in less disturbance than mismatches at the center of the probe. These features, which result in a "smile" shaped graph when plotted, will be found in most plots of single base mismatch after selection of a "correct" core probe, or after accounting for a mismatched probe that is a core probe with respect to another portion of the target sequence. This information will be utilized in either selecting the core probe initially or in checking to ensure that an exactly matched core probe has been selected.

Of course, in certain situations, as noted in in the section above, identification of a core is all that is required such as in, for example, forensic or genetic studies, and the like.

In sequencing studies, this process is then repeated for left and/or right extensions of the core probe. In one example, only right extensions of the core probe are possible. The possible 4-mer extension probes of the core probe are 3'-YXYY and 31-YXYX. Again, the same selection criteria are utilized. Between 31-YXYY and 3'-YXYX, it would normally be found that 31-YXYX would have the strongest binding affinity, and this probe is selected as the correct probe extension. This selection may be confirmed by again plotting the normalized binding affinity of probes with single base mismatches as compared to the core probe.

When a hypothetical plot is illustrated, again, the characteristic "smile" pattern is observed, indicating that the "correct" extension has been selected, i.e., 3'-YXYX. From this information, one would correctly conclude that the sequence of the target is 51-XXYXY.

1.4.4 A Method for Sequencing Genomes

In one embodiment, a method is described for sequencing genomes that is comprised of the steps:

- (1) Obtaining a clone library to be sequenced and mapped;
- (2) Preparing DNA from individual clones in the clone library for comparison experiments;
- (3) Obtaining a long-range probe library relative to the clone library;
- (4) Preparing DNA from members of the long-range probe library for comparison experiments;
- (5) Comparing DNA from the clone library with DNA from the long-range probe library;
- (6) Producing a clone library characterized by long-range probes;
- (7) Obtaining a bin probe library suitable for positioning the DNA sequences of long-range probes relative to the genome;
- (8) Comparing DNA from the bin probe library with DNA from the long-range probe library;
- (9) Producing a long-range probe library whose DNA sequences have been characterized by binning information relative to the genome;
- (10) Combining the clone vs. long-range probe characterization from step 6, together with the long-range probe vs. genome binning characterization from step 9;
- (11) Producing a binning of the clone library;
- (12) Obtaining a short-range probe library relative to the clone library;
- (13) Comparing DNA from the clone library with DNA from the short-range probe library;
- (14) Producing a clone library characterized by short-range probes;
- (15) Combining the long-range binning of the clone library, together with the short-range probing of the clone library from;
- (16) Producing a contig of the clone library which bins and orders clones relative to the genome;
- (17) Forming a tiling path of clones that span genome regions;
- (18) Determining the sequence of said clones, and of the entire genome.

1.4.4.1 Obtaining a clone library to be sequenced and mapped.

The clones may be comprised of large-sized clones that have genomic inserts greater than 250 kb (e.g., YACs), medium-sized clones that have genomic inserts greater than 50 kb, but less than 250 kb (e.g., PACs, BACs, P1s, or YACs), or small-sized clones that have genomic inserts less than 50 kb (e.g., cosmids, plasmids, phage, phagemids, or cDNAs). In the preferred embodiment, the clone library has at least two-fold redundancy relative to the genome. The technology for constructing these clones is well described (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, ed., *Current Protocols in Molecular Biology*. New York, N.Y.: John Wiley and Sons, 1995; N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995; J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning*, Second Edition. Plainview, N.Y.: Cold Spring Harbor Press, 1989), incorporated by reference. Chromosome-specific cosmid clones are available from Los Alamos National Laboratories (Los Alamos, N.Mex.), genome-wide PAC clones from Pieter de Jong (Roswell Park, Buffalo, N.Y.), and the Genethon YAC libraries from the national genome center GESTECs, including the Whitehead Institute (Cambridge, Mass.). Libraries are also provided by commercial vendors, including cDNA libraries (ATCC, Rockville, Md.), P1 libraries (DuPont/Merck Pharmaceuticals, Glenolden, Pa.), BAC libraries (Research Genetics, Huntsville, Ala.), and cDNAs and other genome-wide resources (BIOS Labs, New Haven, Conn.).

1.4.4.1.1 Preparing DNA from individual clones in the clone library for comparison experiments.

In the preferred embodiment, DNA from the clones is prepared for DNA hybridization experiments. For DNA derived from bacterial clones (cosmids, PACs, etc.), two straightforward protocols are: (a) growing up colonies for each clone, and then lysing the bacterial cells to expose the cloned insert DNA, or (b) specifically extracting the DNA material from the clone using DNA prep such as an ion exchange column (Qiagen, Chatsworth, Calif.). When using vectors with more complex genomes (e.g., yeast cells), a species-specific DNA prep (e.g., Alu-PCR or IRE-bubble PCR) is preferred. This DNA from each clone is then gridded onto nylon membranes such as Hybond N+ (Amersham, Arlington Heights, Ill.) to prepare for

subsequent DNA hybridization experiments (Hybond N+ product protocol, ver. 2), incorporated by reference.

1.4.4.1.2 Obtaining a long-range probe library relative to the clone library.

The preferred long-range multiplexed probe is the radiation hybrid (RH) (D. R. Cox, M. Burnmeister, E. R. Price, S. Kim, and R. M. Myers, "Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes," *Science*, vol. 250, pp. 245-250, 1990; S. J. Goss and H. Harris, "New method for mapping genes in human chromosomes," *Nature*, vol. 255, pp. 680-684, 1975; S. J. Goss and H. Harris, "Gene transfer by means of cell fusion: statistical mapping of the human X-chromosome by analysis of radiation-induced gene segregation," *J. Cell. Sci.*, vol. 25, pp. 17-37, 1977), incorporated by reference. Chromosome-specific RH libraries have been constructed for other human chromosomes (M. R. James, C. W. Richard III, J.-J. Schott, C. Yousry, K. Clark, J. Bell, J. Hazan, C. Dubay, A. Vignal, M. Agrapart, T. Imai, Y. Nakamura, M. Polymeropoulos, J. Weissenbach, D. R. Cox, and G. M. Lathrop, "A radiation hybrid map of 506 STS markers spanning human chromosome 11," *Nature Genetics*, vol. 8, no. 1, pp. 70-76, 1994; S. H. Shaw, J. E. W. Farr, B. A. Thiel, T. C. Matise, J. Weissenbach, A. Chakravarti, and C. W. Richard, "A radiation hybrid map of 95 STSs spanning human chromosome 13q," *Genomics*, vol. 27, no. 3, pp. 502-510, 1995; U. Francke, E. Chang, K. Comeau, E.-M. Geigl, J. Giacalone, X. Li, J. Luna, A. Moon, S. Welch, and P. Wilgenbus, "A radiation hybrid map of human chromosome 18," *Cytogenet. Cell Genet.*, vol. 66, pp. 196-213, 1994), incorporated by reference. Whole-genome RHs (WG-RHs) for humans and other mammalian genomes have also been developed (M. A. Walter, D. J. Spillett, P. Thomas, J. Weissenbach, and P. N. Goodfellow, "A method for constructing radiation hybrid maps of whole genomes," *Nature Genet.*, vol. 7, no. 1, pp. 22-28, 1994), incorporated by reference, including the high-energy Stanford set (David Cox, Stanford, Calif.) and the low-energy Genethon set; the DNAs from both WG-RH sets are available (Research Genetics, Huntsville, Ala.).

There are alternative embodiments that can construct long-range multiplexed probes. One alternative embodiment is the use of rare cutter restriction enzymes (e.g., NotI partial digests) to develop large DNA sequences from genomes. These fragments can be purified using pulsed-field gel electrophoresis (D. C. Schwartz and

C. R. Cantor, "Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis," *Cell*, vol. 37, pp. 67-75, 1984), incorporated by reference, and then selectively pooled. A second alternative embodiment is the use of a second clone library that has a larger average insert size than the first clone library in step 1. Subsets of these larger insert clones can be pooled together to form a long-range probe library (relative to the first clone library). A third alternative embodiment which is particularly useful in animal models is the use of genetically inbred strains. With an F1 backcross between strains A and B, the meiotic events produce an interleaving of large chromosomal fragments of strains A and B. A subtractive hybridization can selectively remove the DNA from strain B, leaving behind just the large chromosomal regions of strain A for each backcross individual. This procedure constructs a long-range probe library (relative to the strain A clone library). The subtractive hybridization can be performed by first digesting the backcross individual genome with restriction enzymes, and then using whole genome DNA from strain B bound to solid support to selectively remove the strain B DNA.

1.4.4.1.3 Preparing DNA from members of the long-range probe library for comparison experiments.

The long-range probe DNA often resides in a complex background genome. In the RH embodiment, the background is murine genome, while in the pooled YAC embodiment, the background is the yeast genome. Therefore, the DNA preparations for these long-range probe embodiments preferably use a species-specific DNA extraction and amplification. The particular assay often depends on the clone library used.

When the clonal inserts reside in a complex background genome, such as YACs, inter-Alu hybridization is the preferred approach in step 5. In this case, Alu-PCR preparation of the long-range probes (M. T. Ross and V. P. J. Stanton, "Screening large-insert libraries by hybridization," in *Current Protocols in Human Genetics*, vol. 1, N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed. New York: John Wiley and Sons, 1995, pp. 5.6.1-5.6.34), incorporated by reference, is the preferred embodiment. An alternative embodiment when background hybridization noise may be greater is IRE-bubble PCR (D. J. Munroe, M. Haas, E. Bric, T. Whirton, H. Aburatani, K. Hunter, D. Ward, and D. E. Housman, "IRE-bubble PCR: a rapid method for efficient

and representative amplification of human genomic DNA sequences from complex sources," *Genomics*, vol. 19, no. 3, pp. 506-14, 1994), incorporated by reference.

When the clonal inserts are sufficiently large to contain inter-Alu regions, and the vector genome is not complex (e.g., bacterial), then IRE-bubble PCR is the preferred embodiment. This situation applies to many clone libraries, including cosmids, PACs, BACs, and P1s.

When the clonal inserts are too small to contain inter-Alu subsequences detectable by hybridization (such as cDNAs), an assay that provides for more uniform DNA expression from the long-range probes may be needed. The most preferred embodiment is then to use a multiplicity of restriction enzyme digests, each followed by long PCR between Alu repeats, and to then pool the PCR products to construct a probe. A second approach is a variation on direct selection (M. Lovett, J. Kere, and L. M. Hinton, "Direct selection: a method for the isolation of cDNAs encoded by large genomic regions," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 88, pp. 9628-9632, 1991), incorporated by reference. In this approach, Lovett's cDNAs are replaced by a full restriction digest with a frequent-cutter of the long-range probe DNA, and Lovett's genomic contig is replaced with repetitive DNA (e.g., Alu or Cot-1) that selects for the same genome as the species-specific long-range probe. The result is a PCR amplification (via the end priming sites) of the long-range probe that is species specific (via the Alu selection).

The species-specific DNA is then amplified and labeled for use as a hybridization probe. In the preferred embodiment, this amplification and labeling is performed using a labeled dNTP with the random primer method (A. P. Feinberg and B. Vogelstein, "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity," *Analyt. Biochem.*, vol. 132, pp. 6-13, 1983; N.J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995), incorporated by reference. In one embodiment, ³²P- dNTP is incorporated into a random primer PCR amplification, possibly using a kit such as the DECprime II DNA labeling kit (Ambion, Austin, Tex.). Other isotopes such as ³⁵S or ³³P can be used. In alternative embodiments, nonisotopic labeling is performed (L. J. Kricka, ed., *Nonisotopic Probing, Blotting, and Sequencing*, Second Edition. San Diego, Calif.: Academic Press, 1995), incorporated by reference.

1.4.4.1.4 Comparing DNA from the clone library with DNA from the long-range probe library.

The labeled long-range probe DNA is hybridized against the gridded clone library (A. P. Monaco, V. M. S. Lam, G. Zehetner, G. G. Lennon, C. Douglas, D. Nizetic, P. N. Goodfellow, and H. Lehrach, "Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products," *Nucleic Acids Res.*, vol. 19, no. 12, pp. 3315-3318, 1991), incorporated by reference. In an alternative embodiment, the roles of the long-range probe library and the clone library are reversed, with the long-range probe immobilized on the membrane and the label on the clone.

The hybridization comparison is done by preannealing the probe with 25 ng of Cot-1 DNA (Gibco-BRL, Grand Island, N.Y.) for 2 hours at 37°C. before adding to the prehybridization mix. The nylon filters containing the spotted clone DNA is then prehybridized overnight per manufacturer's instructions (Amersham, Arlington Heights, Ill.), except for the addition of sheared, denatured human placental DNA at a final concentration of 50 ng/ml. Filters are hybridized overnight at 68°C., washed three times with final wash of 0.1 SSPE/0.1% SDS at 72° C., before exposing to autoradiographic film for 1 to 8 days. The exposed film image is then electronically scanned into a computer with memory. A phosphorimager (Molecular Dynamics, Sunnyvale, Calif.) or other electronic device can be used for imaging without the use of film.

For every RH hybridization probing, each of the clone positions on the autoradiographs of the gridded filters are scored on a numerical scale, such as 1-5, with 1 negative, 2 equivocal., 3 weakly positive, 4 positive, and 5 strongly positive. When duplicate typings are available, the maximum of the two scores is used, since there is a very high false-negative rate in the hybridization data. This data entry can be facilitated by use of an interactive computer program that presents the electronic image of the filter on a computer display, or by automated computer interpretation of the scanned image.

1.4.4.2 Producing a clone library characterized by long-range probes.

The hybridization experiments construct a table of scores that compare the DNA from clones against DNA from long-range probes for detectable sequence similarity, and thus presumed genomic colocalization. The scores are rescaled so that

the new scaling is approximately linear (C. C. Clogg and E. S. Shihadeh, *Statistical Models for Ordinal Variables*. Thousand Oaks, Calif.: Sage Press, 1994), incorporated by reference. That is, a unit increase in the scaling indicates a unit increase in the confidence one holds that the clone actually hybridized with the long-range probe. An equivocal event is scored as a 0, since it was equally likely to be negative or positive. A negative event is scored as -1, since there is high confidence that no observable hybridization has occurred; both positive and strongly positive events are scored as 1, since there is certainty that a hybridization event has occurred. A weakly positive event can be scored at 0.67 when a single typing is available, since there is considerably more confidence that it is positive than negative, and is considered equivocal when duplicate typings were available. For any scale used, the data is scored in a manner determined by the laboratory investigator and data analyst. This rescaled clone vs. probe comparison table A is stored in the memory of a computational device.

With perfectly clean comparison data (i.e., very low false negative and false positive rates), this table A might suffice for ordering the clones using conventional RH mapping methods. However, the high-throughput hybridization experiments incur a large noise cost. Therefore, some correction data is required to accurately map the clones. This correction stage is performed in the following steps.

1.4.4.2.1 Obtaining a bin probe library suitable for positioning the DNA sequences of long-range probes relative to the genome.

In the preferred embodiment, the bin probe library is comprised of sequence-tagged sites (STSs). For positional cloning applications, many of the STSs are preferably made polymorphic. The genetic or physical markers to be used for each STS are obtained as PCR primer sequences pairs and PCR reaction conditions from available Internet databases (Genbank, Bethesda, Md.; GDB, Baltimore, Md.; EMBL, Cambridge, UK; Genethon, Evry, France; Stanford Genome Center, Stanford, Calif.; Whitehead Institute Genome Center, Cambridge, MA; G. Gyapay, J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop, and J. Weissenbach, "The 1993-94 Genethon Human Genetic Linkage Map," *Nature Genetics*, vol. 7, no. 2, pp. 246-339, 1994; Hilliard, Davison, Doolittle, and Roderick, Jackson laboratory mouse genome database, Bar Harbor, Me.; MapPairs, Research Genetics, Huntsville, Ala.), incorporated by reference. Alternatively, STSs can be

constructed using existing techniques (Sambrook, J., Fritsch, E. F., and Manjaris, T. 1989. *Molecular Cloning*, second edition. Plainview, N.Y.: Cold Spring Harbor Press; N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995), incorporated by reference.

In a first alternative embodiment, the locations of the long-range probe fragments are localized on the genome by fluorescence in situ hybridization (FISH) studies. In these FISH studies, the nuclear DNA of the genome serves as the bin probe. In a second alternative embodiment, the binning is effected by comparison with previously positioned DNA probes, including mapped clone libraries, ESTs, or PCR primers.

1.4.4.2.2 Comparing DNA from the bin probe library with DNA from the long-range probe library.

In the preferred embodiment, PCR amplifications are carried out between the STSs in the bin probe library and the RH (or other) DNAs in the long-range probe library. Subsequent detection for presence or absence of PCR products (+/- scores) is carried out either by gel electrophoresis or by internal oligonucleotide hybridizations.

The orders of the STSs relative to the genome are then determined using computational or statistical methods (M. Boehnke, "Radiation hybrid mapping by minimization of the number of obligate chromosome breaks," *Genetic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes*. Cytogenet Cell Genet, vol. 59, pp. 96-98, 1992; M. Boehnke, K. Lange, and D. R. Cox, "Statistical methods for multipoint radiation hybrid mapping," *Am. J. Hum. Genet.*, vol. 49, pp. 1174-1188, 1991; A. Chakravarti and J. E. Reefer, "A theory for radiation hybrid (Goss-Harris) mapping: application to proximal 21q markers," *Generic Analysis Workshop 7: Issues in Gene Mapping and the Detection of Major Genes*. Cytogenet Cell Genet, vol. 59, pp. 99-101, 1992), incorporated by reference. Physical distances are then computed using maximum likelihood estimation.

In the first alternative FISH embodiment of step 7, DNA from the long-range probes (e.g., species-specific PCR products) are fluorescently labeled, and then hybridized back onto the genome. The fragment positions on the genome of the probes are then visualized using fluorescent microscopic imaging. Linear fractional length measurements on the metaphase spreads of chromosomes are then performed

to determine the bin positions of the fragments. In the second alternative embodiment of step 7, DNA from the previously positioned bin probes is hybridized to DNA from the long-range probes.

Detailed protocols for these methods have been described (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, ed., *Current Protocols in Molecular Biology*. New York, N.Y.: John Wiley and Sons, 1995; N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995), incorporated by reference.

1.4.4.2.3. Producing a long-range probe library whose DNA sequences have been characterized by binning information relative to the genome.

The procedures produce a data table which compares the DNA content of the long-range probes to bins on the genome. In the preferred embodiment, this is a table B of long-range probes (the rows of B) vs. ordered STSs (the columns of B). The pairwise distance information between the ordered STSs is also recorded. In alternative embodiments, the table can be arranged similarly.

Knowledge of the genomic positions of the RH fragments enables the desired correction of noisy RH hybridization data, as described next.

1.4.4.3 Producing a binning of the clone library.

The procedures of step 10 produce a table which bins each clone relative to the genome. In the preferred embodiment, this is a table C of clones (the rows of C) vs. ordered bins (the columns of C). Each entry in the table describes the confidence that the clone is located in the bin.

Note that this result C is a binning of clones, not a contig. To form the desired set of mapped overlapping clones, a short-range probing is preferably performed. This probing and contig formation is performed in the following steps.

1.4.4.3.1 Obtaining a short-range probe library relative to the clone library.

Since current clone mapping technology is based on short-range probing, there is a large number of workable approaches. The preferred embodiment uses hybridization assays based on oligonucleotide probes. The design of such experiments has been described (A. J. Cuticchia, J. Arnold, and W. E. Timberlake, "PCAP: probe choice and analysis package, a set of programs to aid in choosing synthetic oligomers

for contig mapping," CABIOS, vol. 9, no. 2, pp. 201-203, 1992; Y.-X. Fu, E. W. Timberlake, and J. Arnold, "On the design of genome mapping experiments using short synthetic oligonucleotides," Biometrics, vol. 48, pp. 337-359, 1992; H. Lehrach, A. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A. P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka, "Hybridization fingerprinting in genome mapping and sequencing," in Genetic and Physical Mapping I: Genome Analysis, K. E. Davies and S. M. Tilghman, ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory, 1990, pp. 39- 81; A. Poustka, T. Pohl, D. P. Barlow, G. Zehetner, A. Craig, F. Michiels, E. Erlich, A.-M. Frischauf, and H. Lehrach, "Molecular approaches to mammalian genetics," in Cold Spring Harbor Symp. Quant. Biol., vol. 51. 1986, pp. 131-139), incorporated by reference.

An efficient design produces 25 to 200 small (preferably 5 bp-15 bp) oligonucleotides which each hybridize, on average, to 5%-95% of the clones. The oligonucleotide sequences are generally designed to preferentially detect sequences that are related to the genes in the genome, rather than to repetitive elements in the genome or to the cloning vector. This selective bias can be achieved either by experimental probings, or by examination of the sequences to be compared. Once designed, these oligonucleotides are preferably ordered from a DNA synthesis service (Research Genetics, Huntsville, Ala.). Alternatively, they can be synthesized on a DNA synthesizer (Applied Biosystems, Foster City, Calif.).

Alternative hybridization embodiments include using clones (or their PCR products) to probe clone libraries, using pools of clones as hybridization probes, and using Southern blotting of digested clones with repetitive element hybridization probes. Enzymatic methods include gel electrophoresis of restriction endonuclease digests of clones, PCR-based STS comparisons, and hybrid methods such as Alu fingerprinting. Other short-range probes can be formed by selective or random retention of fragments produced by genome cutting.

For experimental efficiency, many of these short-range probes work in a multiplexed way, and probe one or more genome regions simultaneously. These probes include oligonucleotides, pooled clones, and repetitive-element fingerprint probes.

1.4.4.3.2 Comparing DNA from the clone library with DNA from the short-range probe library.

This is done by comparison experiments using standard protocols. In the preferred embodiment, DNA from the clones in the clone library is spotted onto nylon membranes. This DNA is comprised of lysed colonies, DNA preps, or species-specific PCR products. The membranes are then prepared for hybridization. Each oligonucleotide short-range probe is then labeled, preferably with ^{32}P using a kinase. The labeled probe is then hybridized to the membranes, followed by rinsing, stringent washing, and autoradiography. The filters may be stripped for subsequent reuse. The autoradiograph spots are then scored on a binary or more continuous (e.g., 0-255) scale.

Specific oligonucleotide hybridization protocols for particular clone libraries and oligonucleotides have been described (A. G. Craig, D. Nizetic, J. D. Hoheisel, G. Zehetner, and H. Lehrach, "Ordering of cosmid clones covering the herpes simplex virus type I," *Nucleic Acids Res.*, vol. 18, no. 9, 2653-60, 1990; R. Drmanac, Z. Strezoska, I. Labat, S. Drmanac, and R. Crkvenjakov, "Reliable hybridization of oligonucleotides as short as six nucleotides," *DNA Cell Biol.*, vol. 9, no. 7, pp. 527-534, 1990; J. D. Hoheisel, G. G. Lennon, G. Zehetner, and J. Lehrach, "Use of high coverage reference libraries of *Drosophila melanogaster* for relational analysis," *J. Mol. Biol.*, vol. 220, pp. 903-914, 1991; F. Michiels, A. G. Craig, G. Zehetner, G. P. Smith, and H. Lehrach, "Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries," *CABIOS*, vol. 3, pp. 203-210, 1987; D. Nizetic, R. Drmanac, and J. Lehrach, "An improved bacterial colony lysis procedure enables direct DNA hybridization using short (10, 11 bases) oligonucleotides to cosmids," *Nucleic Acids Res.*, vol. 19, pp. 182, 1991), incorporated by reference.

For alternative short-range probes, the comparison protocols are described (see cited references above).

1.4.4.3.3 Producing a clone library characterized by short-range probes.

The comparison experiments of the previous step construct a table D of scores that compare the DNA from clones against DNA from short-range probes. These provide measures of genomic colocalization and distance.

In this step, or in the following step 15, contigs can be formed from the short-range characterization data of the clones. In the preferred embodiment, each clone's score signature relative to the oligonucleotides is compared against other clones' score

signatures. Pairs of clones having similar score signatures are inferred to be close, and their distances can be estimated. The preferred ordering method is simulated annealing (W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1988), incorporated by reference. Effective contiging algorithms have been described (A. J. Cuticchia, J. Arnold, and W. E. Timberlake, "ODS: ordering DNA sequences, a physical mapping algorithm based on simulated annealing," *CABIOS*, vol. 9, no. 2, pp. 215-219, 1992; A. J. Cuticchia, J. Arnold, and W. E. Timberlake, "The Use of Simulated Annealing in Chromosome Reconstruction Experiments Based on Binary Scoring," *Genetics*, vol. 132, pp. 591-601, 1992; A. Milosavljevic, Z. Strezoska, M. Zeremski, D. Grujic, T. Paunesku, and R. Crkvenjakov, "Clone clustering by hybridization," *Genomics*, vol. 27, no. 1, pp. 83-89, 1995), incorporated by reference.

For alternative short-range probes, the contiging analysis procedures use analogous comparison data and search procedures, and have been described (D. O. Nelson and T. P. Speed, "Statistical issues in constructing high resolution physical maps," *Statistical Science*, vol. 9, no. 3, pp. 334-354, 1994; E. Branscomb, T. Slezak, R. Pae, D. Galas, and al., "Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries," *Genomics*, vol. 8, pp. 351-366, 1990; S. G. Fisher, E. Cayanis, J. J. Russo, I. Sunjevaric, B. Boukhgalter, P. Zhang, M.-T. Yu, R. Rothstein, D. Warburton, I. S. Edelman, and A. Efstratiadis, "Assembly of ordered contigs of cosmids selected with YACs of human chromosome 13," *Genomics*, vol. 21, pp. 525-537, 1994; R. Mort, A. Grigoriev, E. Maier, J. Hoheisel, and H. Lehrach, "Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*," *Nucleic Acids Research*, vol. 21, no. 8, pp. 1965-1974, 1993), incorporated by reference.

1.4.4.3.4 Forming a tiling path of clones that span genome regions.

From an accurate clone map of a genome, a (not necessarily unique) subset of clones that cover the genome can be identified. This identification is done by starting from a leftmost clone by moving rightward from a selected clone A, selecting a neighbor B which overlaps A, and then iteratively continuing from B. A constraint

can be placed on this process to find tiling paths having small or minimal length, where length is defined as the sum of the insert sizes of the component clones.

In the preferred embodiment, (minimal) tiling paths have immediately utility for finding genes. This is because the inner product map integrates genetic markers (polymorphic STSs) together with the clones that fully cover the genome region containing the gene of interest. This considerably reduces the search effort for cloning the gene. Even greater utility for positional/candidate cloning (F. S. Collins, "Positional cloning moves from perditional to traditional.," *Nature Genet.*, vol. 9, no. 4, pp. 347-350, 1995), incorporated by reference, is present when a map of ESTs, expressed cDNAs, or exons is also integrated into the map.

1.4.4.3.5 Determining the sequence of said clones, and of the entire genome.

In the preferred embodiment, each mapped clone is selected in turn from a minimum tiling path. This clone is then subcloned into M13 sequencing vectors. For each M13 subclone, nested deletions are constructed for use in DNA sequencing. For each deletion clone, a DNA sequencing template is prepared. This template is then sequenced by the dideoxy method, preferably using an automated DNA sequencer, such as an A. L. F. (Pharmacia Biotech, Piscataway, N.J.) or an ABI/373 or ABI/377 (Applied Biosystems, Foster City, Calif.), and 100-500 bp of sequence determined. In addition to this "shotgun" phase, in which an initial read is taken from each subclone using a universal primer, a "walking" phase takes additional reads from selected subclones by use of custom primers. Complete protocols for these and related sequencing steps have been described (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, ed., *Current Protocols in Molecular Biology*. New York, N.Y.: John Wiley and Sons, 1995; N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995).

The sequences of the nested deletion clones are assembled into the complete sequence of the subclone by matching overlaps. The subclone sequences are then assembled into the sequence of the mapped clone. The sequences of the mapped clones are assembled into the complete sequence of the genome by matching overlaps. Computer programs are available for these tasks (Rodger Staden programs, Cambridge, UK; DNASTar, Madison, Wis.). Following sequence assembly, current analysis practice includes similarity and homology searches relative to sequence

databases (Genbank, Bethesda, Md.; EMBL, Cambridge, UK; Phil Green's GENEFINDER, Seattle, Wash.) to identify genes and repetitive elements, infer function, and determine the sequence's relation to other parts of the genome and cell.

1.4.4.4.6 Application of Strategies

Such strategies have been successfully applied to sequencing the genomes of several bacteria (Human Genome Sciences, Gaithersburg, Md.), including *E. coli* (G. Plunkerr and al., "Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes," *Nucl. Acids Res.*, vol. 21, pp. 3391-3398, 1993), incorporated by reference, and higher organisms, including yeast (S. G. Oliver and al., "The complete sequence of yeast chromosome III," *Nature*, vol. 357, pp. 38-46, 1992), incorporated by reference, human (A. Martin-Gallardo and al., "Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3," *Nature Genet.*, vol. 1, pp. 34-39, 1992), incorporated by reference, mouse (R. K. Wilson and al., "Nucleotide sequence analysis of 95 kb near the 3' end the murine T-cell receptor alpha/delta chain locus: strategy and methodology," *Genomics*, vol. 13, pp. 1198-1208, 1992), incorporated by reference, and *C. elegans* (R. Wilson and al., "2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*," *Nature*, vol. 368, pp. 32-38, 1994; J. Sulston, Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, S. Dear, A. Coulson, M. Craxton, M. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough, and R. Waterston, "The *C. elegans* genome sequencing project: a beginning," *Nature*, vol. 356, pp. 37-41, 1992), incorporated by reference. The automated sequencing of large genome regions from mapped cosmid (or other) clones is now routine in several centers (Sanger Center, Cambridge, UK; Washington University, St. Louis, Mo.), with very low error at an average cost of \$0.50 or less per base. Specific strategies and protocols for these efforts have been detailed (H. G. Griffin and A. M. Griffin, ed., *DNA Sequencing: Laboratory Protocols*. New Jersey: Humana, 1992), incorporated by reference.

The current best mode for sequencing is gel electrophoresis on polyacrylamide gels, possibly using fluorescence detection. Newer technologies for DNA size separation are being developed that are applicable to DNA sequencing, including ultrathin gel slabs (A. J. Kostichka, M. L. Marchbanks, R. L. Brumley Jr., H. Drossman, and L. M. Smith, "High speed automated DNA sequencing in ultrathin

slab gels," *Bio/Technology*, vol. 10, pp. 78-81, 1992), incorporated by reference, capillary arrays (R. A. Mathies and X. C. Huang, "Capillary array electrophoresis: an approach to high-speed, high-throughput DNA sequencing," *Nature*, vol. 359, pp. 167-169, 1992), incorporated by reference, and mass spectrometry (K. J. Wu, A. Stedding, and C. H. Becker, "Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix," *Rapid Commun. Mass Spectrom.*, vol. 7, pp. 142-146, 1993), incorporated by reference. DNA sequencing without the use of gel electrophoresis has also been done using sequencing by hybridization methodologies (R. Drmanac, S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zereński, J. Snoddy, W. K. Funkhouser, B. Koop, and L. Hood, "DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing," *Science*, vol. 260, pp. 1649-1652, 1993; E. M. Southern, U. Maskos, and J. K. Elder, "Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models," *Genomics*, vol. 13, pp. 1008-10017, 1991; S. P. A. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed spatially addressable parallel chemical synthesis," *Science*, vol. 251, pp. 767-773, 1991), incorporated by reference. Another approach is base addition sequencing strategy (BASS), which uses synchronized DNA polymer construction to determine the sequence of unknown DNA templates (P. C. Cheeseman, "Method for sequencing polynucleotides," U.S. Pat. No. 5,302,509; filed Feb. 27, 1991, published Apr. 12, 1994; A. Rosenthal, K. Close, and S. Brenner, "DNA sequencing method," Patent #PCT WO 93/21340; filed Apr. 22, 1992, published Oct. 28, 1993; R. Y. Tsien, P. Ross, M. Fahrenstock, and A. J. Johnston, "DNA sequencing," Patent #PCT WO 91/06678; filed Oct. 26, 1990, published May 16, 1991), incorporated by reference.

1.4.5 Insertion of a Genomic Fragment into an Appropriate Host Vector

In another embodiment, the process begins with a fragment of DNA, such as a genomic fragment, which is inserted into an appropriate host vector capable of accommodating it. For example, a BAC vector can accommodate approximately 140 kb of DNA; a cosmid vector can accommodate approximately 40 kb. A composition comprised of these insert-containing vectors is randomly sheared using standard methods, such as sonication, to obtain fragments suitable for transposon-based sequencing--i.e., about 2-5 kb, preferably 3-4 kb, on the average.

The resulting subfragments are ligated into cloning vectors to create a first library of subclones representing the original fragment. Because the subclones in this library will be used as target plasmids for transposon-mediated sequencing, the size of the cloning vector should be minimized; preferably it should contain only a selectable marker, an origin of replication, and an insertion site. A suitable host plasmid is pOT2; the subfragments obtained by shearing the original composition are end-repaired, ligated to suitable restriction site containing adapters, and inserted into the host vector. Suitable adapters for the pOT2 vector contain BstXI sites.

The resulting cloning vectors with their inserts are then transfected into bacteria, typically *E. coli*, for clonal growth. This first library should contain a 15-20-fold representation of the original fragment of DNA. For example, if the original fragment is approximately 40 kb, and the subclones contain inserts of approximately 4 kb, 200 such clones would be required for a 20-fold representation of the original fragment.

1.4.5.1 Hybridization Screening

As pointed out above, this first library will contain subclones which do not contain DNA derived from the original fragment to be sequenced. In order to eliminate these subclones, a preliminary hybridization screen is conducted. The required number of subclones is prepared for hybridization screening, for example, by plating in 96-well plates and transferring to filters. The filters are then probed with the original fragment insert to weed out any colonies which do not contain DNA which represents portions of the original fragment. This checks the quality of the library and eliminates subclones that contain only host cloning vector for the original fragment or contaminating bacterial DNA.

1.4.5.2 2nd Library Formation by Subclones that Contain Inserts

The subclones confirmed to contain inserts derived from the fragment to be sequenced form a second library. The number of subclones in this library should be sufficient to contain a 7-8x times. representation of the fragment. Each subclone is individually sequenced from one end of the insert. This is straightforward, since the sequence information in the cloning vector provides sufficient information to design appropriate primers. Typically, about 400-450 nucleotides into the insert is read. In addition to the requirement for 7-8x times. coverage of the fragment when the complete insert sequences of the subclones are obtained, there must be sufficient sequence information available from this end sequencing to represent a 1x times. coverage of the fragment. Thus, if the original fragment contained 40 kb and 400 nucleotides into the insert is read, 100 clones would be required. The resulting sequence information is organized into a computer-readable form for searching. A DNA sequence comparison algorithm can be used for subsequent comparisons, such as the NCBI program BLASTN.

The criteria used to determine the number of subclones used to establish the database in the method described above are that low sequencing redundancy must be maintained and a complete path must be available within the set of subclones chosen to provide complete coverage of the original fragment. In addition, the number must be chosen so that there is a high probability of finding the next subclone when searching with the newly sequenced end sequence.

A method similar to that employed by Chen, E. et al. Genomics (1993) 17:651-666, is used. Lander and Waterman (cite) conclude that the maximum number of sequence islands occurs at $C=(1-\theta)^{-1}$, where C is the sequence coverage and theta is the ratio of the number of bases required to detect the true overlap to the sequence read length. As theta approaches zero, sequence coverage of 1 will produce the maximum number of sequence islands. In order to achieve the highest efficiency database, enough end-sequence data should be generated to obtain about 1x times. coverage.

In addition, the subclone coverage--i.e., the redundancy based on the complete sequence contained in the number of subclones chosen--is important. A subclone coverage factor of 7x-8x times provides a 99.9% probability that each nucleotide in the fragment will actually reside in the library. This requires only about 100 subclones averaging 3 kb in size for a 40 kb fragment.

Sequence information from the host vector for the original fragment is used as the first query and reveals which subclones in the library are hybrid vector/fragment insert subclones. These will identify the two ends of the original fragment. One subclone representing each end, preferably that containing the least amount of vector sequence, is selected for further sequencing. The insert of the identified subclone will be sequenced from the opposite end from that previously sequenced-- i.e., opposite the end containing the vector sequence. The new sequence information (which is now derived from the fragment) is used as the next query. This identifies additional subclones which contain additional nucleotide sequence farther in from the end of the original fragment. The next identified subclone is then also sequenced from the opposite end of the insert from that used to place it in the database and the new sequence information used as the next query. The process is continued sequentially until a subclone path through the fragment is obtained. The subclone path will represent the collection of subclones which completely define the fragment from which they originated, and their correct relative positions are known.

At any point in this process, if there are no responses to the query, additional sequence can be obtained from the subclones already identified and this sequence used as the query.

Once the subclone path is determined, it remains only to complete the sequencing of the subclones involved in the path. According to the method of the invention, this is accomplished using the transposon-mediated method of Strathmann incorporated by reference hereinabove. Use of this method to complete the sequence information for the fragment has been designated "minimal assembled path" (MAP) sequencing. The name is apt because the information provided by the subclone path can be used to determine the minimal sequencing path through the identified subclones. For example, if two subclones overlap over 1 kb, transposon insertions can be selected so that the overlap region is sequenced only once. Thus, although theoretically each of the subclones obtained to define the path can be completely sequenced using the transposon-mediated method, only sufficient portions of these subclones need be sequenced to obtain the complete sequence of the original fragment.

1.4.6 Methods of Determining A Nucleic Acid Sequence through Enzymatic Sequencing

In another embodiment, improved methods of determining a nucleic acid sequence through enzymatic sequencing are provided. In the subject methods, primers are used in combination with capturable chain terminators to produce primer extension products capable of being captured on a solid phase, where the primer extension products may be labeled, e. g. by employing labeled primers to generate the primer extension products. Following generation of the primer extension products, the primer extension products are isolated through capture on a solid phase. The isolated primer extension products are then released from the solid phase, size separated and detected to yield sequencing data from which the nucleic acid sequence is determined.

Methods of determining the sequence of a nucleic acid, e.g. DNA, by enzymatic sequencing are well known in the art and described in Sambrook et al., *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 1989) and Griffin and Griffin, "DNA Sequencings, Recent Innovations and Future Trends," *Applied Biochemistry and Biotechnology* (1993) 38: 147-159, the disclosures of which are herein incorporated by reference. The Sanger method is shown schematically herein. Generally, in enzymatic sequencing methods, which are also referred to as Sanger dideoxy or chain termination methods, differently sized oligonucleotide fragments representing termination at each of the bases of the template DNA are enzymatically produced and then size separated yielding sequencing data from which the sequence of the nucleic acid is determined. The results of such size separations are shown herein. The first step in such methods is to produce a family of differently sized oligonucleotides for each of the different bases in the nucleic acid to be sequenced, e.g. for a strand of DNA comprising all four bases (A, G, C, and T) four families of differently sized oligonucleotides are produced, one for each base. To produce the family of differently sized oligonucleotides, each base in the sequenced nucleic acid, i.e. template nucleic acid, is combined with an oligonucleotide primer, a polymerase, nucleotides and a dideoxynucleotide corresponding to one of the bases in the template nucleic acid. Each of the families of oligonucleotides are then size separated, e.g. by electrophoresis, and detected to obtain sequencing data, e.g. a separation pattern or electropherogram, from which the nucleic acid sequence is determined.

Before further describing the subject methods in greater detail, the critical chain terminator reagents employed in the subject methods will be discussed. Critical to the subject methods is the use of capturable chain terminators to produce the families of different sized oligonucleotide fragments (hereinafter referred to as primer extension products) comprising a capture moiety at the 3' terminus. The primer sequences employed to generate the primer extension products will be sufficiently long to hybridize the nucleic acid comprising the target or template nucleic acid under chain extension conditions, where the length of the primer will generally range from 6 to 40, usually 15 to 30 nucleotides in length. The primer will generally be a synthetic oligonucleotide, analogue or mimetic thereof, e.g. a peptide nucleic acid. Although the primer may hybridize directly to the 3' terminus of the target nucleic acid where a sufficient portion of this terminus of the target nucleic acid is known, conveniently a universal primer may be employed which anneals to a known vector sequence flanking the target sequence. Universal primers which are known in the art and commercially available include pUC/M 13, g t10, g t11 and the like.

1.4.6.1 Primers Comprise a Detectable Label

In one preferred embodiment of the subject invention, the primers employed in the subject invention will comprise a detectable label. A variety of labels are known in the art and suitable for use in the subject invention, including radioisotopic, chemiluminescent and fluorescent labels. As the subject methods are particularly suited for use with methods employing automated detection of primer extension products, fluorescent labels are preferred. Fluorescently labeled primers employed in the subject methods will generally comprise at least one fluorescent moiety stably attached to one of the bases of the oligonucleotide.

The primers employed in the subject invention may be labeled with a variety of different fluorescent moieties, where the fluorescer or fluorophore should have a high molar absorbance, where the molar absorbance will generally be at least $10^4 \text{ cm}^{-1} \text{ M}^{-1}$, usually at least $10^4 \text{ cm}^{-1} \text{ M}^{-1}$ and preferably at least $10^5 \text{ cm}^{-1} \text{ M}^{-1}$, and a high fluorescence quantum yield, where the fluorescence quantum yield will generally be at least about 0.1, usually at least about 0.2 and preferably at least about 0.5.

For primers labeled with a single fluorescer, the wavelength of light absorbed by the fluorescer will generally range from about 300 to 900 nm, usually from about 400 to 800 nm, where the absorbance maximum will typically occur at a wavelength

ranging from about 500 to 800 nm. Specific fluorsceners of interest for use in singly labeled primers include: fluorescein, rhodamine, BODIPY, cyanine dyes and the like, and are farther described in Smith et al., *Nature* (1986) 321: 647-679, the disclosure of which is herein incorporated by reference.

Of particular interest for use in the subject methods are energy transfer labeled fluorescent primers, in which the primer comprises both a donor and acceptor fluorescer component in energy transfer relationship. Energy transfer labeled primers are described in PCT/US95/01205 and PCT/US96/13134, as well as in Ju et al., *Nature Medicine* (1996)2:246-249, the disclosures of which are herein incorporated by reference.

In an alternative embodiment of the subject invention, instead of using labeled primers labeled deoxynucleotides are employed, such as fluorescently labeled dUTP, which are incorporated into the primer extension product resulting in a labeled primer extension product.

The dideoxynucleotides employed as capturable chain terminators in the subject methods will comprise a functionality capable of binding to a functionality present on a solid phase. The bond arising from reaction of the two functionalities should be sufficiently strong so as to be stable under washing conditions and yet be readily disruptable by specific chemical or physical means. Generally, the chain terminator dideoxynucleotide will comprise a member of a specific binding pair which is capable of specifically binding to the other member of the specific binding pair present on the solid phase. Specific binding pairs of interest include ligands and receptors, such as antibodies and antigens, biotin and strept/avidin, sulfide and gold (Cheng & Brajter-Toth, *Anal.Chem.* (1996)68:4180-4185, and the like, where either the ligand or the receptor, but usually the ligand, member of the pair will be present on the chain terminator. Of particular interest for use as chain terminators are biotinylated dideoxynucleotides, where such dideoxynucleotides are known in the art and available commercially, e. g. biotin- 11 -ddATP, biotin- 11 -ddGTP, biotin- 11 -ddCTP and biotin- 11 -ddTTP, and the like.

1.4.6.2 Subject Methods

Turning now to the subject methods, the nucleic acids which are capable of being sequenced by the subject methods are generally deoxyribonucleic acids that have been cloned in appropriate vector, where a variety of vectors are known in the

art and commercially available, and include M 13mp 18, pGEM, pSport and the like. The first step in the subject method is to prepare a reaction mixture for each of the four different bases of the sequence to be sequenced or target DNA. Each of the reaction mixtures comprises an enzymatically generated family of primer extension products, usually labeled primer extension products, terminating in the same base. In other words, in practicing the subject method, one will first generate an "A", "G", "C", and "T," family of differently sized primer extension products using the target DNA as template. To generate the four families of differently sized primer extension products, template DNA, a DNA polymerase, primer (which may be labeled), the four different deoxynucleotides, and capturable dideoxynucleotides are combined in a primer extension reaction mixture. The components are reacted under conditions sufficient to produce primer extension products which are differently sized due to the random incorporation of the capturable dideoxynucleotide and subsequent chain termination. Thus, to generate the "A" family of differently sized primer extension products, the above listed reagents will be combined into a reaction mixture, where the dideoxynucleotide is ddATP modified to comprise a capturable moiety, e.g. biotinylated ddATP, such as biotin-11-ddATP. The remaining "G", "C", and "T" families of differently sized primer extension products will be generated in an analogous manner using the appropriate dideoxynucleotide.

Where labeled primers are employed to generate each of the families of primer extension products, the labeled primers may be the same or different. Preferably, the labeled primer employed will be different for production of each of the four families of primer extension products, where the labels will be capable of being excited at substantially the same wavelength and yet will provide a distinguishable signal. The use of labels with distinguishable signals affords the opportunity of separating the differently sized primer extension products when such products are together in the same separation medium. This results in superior sequencing data and therefore more accurate sequence determination. For example, one can prepare the "A" family of primer extension products with a first fluorescent label capable of excitation at a wavelength from about 470 to 480 nm which fluoresces at 525 nm. The label used in production of "G", "C", and "T" families will be excitable at the same wavelength as that used in the "A" family, but will emit at 555 nm, 580 nm, and 605 nm respectively. Accordingly, the primer extension labels are designed so that all four of

the labels absorb at substantially the same wavelength but emit at different wavelengths, where the wavelengths of the emitted light differ in detectable and differentiable amounts, e.g. differ by at least 15 nm. The next step in the subject method is isolation of the primer extension products. The primer extension products are isolated by first capturing the primer extension products on a solid phase through the capture moiety at the 3' terminus of the primer extension product and then separating the solid phase from the remaining components of the reaction mixture.

Capture of the primer extension products occurs by contacting the reaction mixture comprising the family of primer extension products with a solid phase. The solid phase has a member of a specific binding pair on its surface. The other member of the specific binding pair is bonded to the primer extension products, as described above. Contact will occur under conditions sufficient to provide for stable binding of the specific binding pair members. A variety of different solid-phases are suitable for use in the subject methods, such phases being known in the art and commercially available. Specific solid phases of interest include polystyrene pegs, sheets, beads, magnetic beads, gold surface and the like. The surfaces of such solid phases have been modified to comprise the specific binding pair member, e.g. for biotinylated primer extension products, streptavidin coated magnetic bead may be employed as the solid phase.

Following capture of the primer extension reaction products on the solid phase, the solid phase is then separated from the remaining components of the reaction mixture, such as template DNA, excess primer, excess deoxy- and dideoxynucleotides, polymerase, salts, extension products which do not have the capture moiety, and the like. Separation can be accomplished using any convenient methodology. The methodology will typically comprise washing the solid phase, where further steps can include centrifugation, and the like. The particular method employed to separate the solid-phase is not critical to the subject invention, as long as the method employed does not disrupt the bond linking the primer extension reaction product from the solid-phase.

The primer extension products are then released from the solid phase. The products may be released using any convenient means, including both chemical and physical means, depending on the nature of the bond between the specific binding pair members. For example, where the bond is a biotin-streptavidin bond, the bond may be

disrupted by contacting the solid phase with a chemical disruption agent, such as formamide, and the like, which disrupts the biotin-streptavidin bond and thereby releases the primer extension product from the solid phase. The released primer extension products are then separated from the solid phase using any convenient means, including elution, centrifugation and the like.

The next step in the subject method is to size separate the primer extension products. Size separation of the primer extension products will generally be accomplished through electrophoresis, in which the primer extension products are moved through a separation medium under the influence of an electric field applied to the medium, as is known in the art. Alternatively, for sequencing with Mass Spectrometry (MS) where unlabeled primer extension products are detected, the sequencing fragments are separated by the time of the flight chamber and detected by the mass of the fragments. See Roskey et al., Proc. Natl. Acad. Sci. USA (1996) 93: 4724-4729. The subject methodology is especially important for obtaining accurate sequencing data with MS, because the subject methodology offers a means to load only the primer extension products terminated with the capturable chain terminators, eliminating all other masses,,thereby producing accurate results.

In methods in which the fragments are size separated, the size separated primer extension products are then detected, where detection of the size separated products yields sequencing data from which the sequence of the target or template DNA is determined. For example, where the families of fragments are separated in a traditional slab gel in four separate lanes, one corresponding to each base of the target DNA, sequencing data in the form of a separation pattern is obtained. From the separation pattern, the target DNA sequence is then determined, e.g. by reading up the gel. Alternatively, where automated detectors are employed and all of the reaction products are separated in the same electrophoretic medium, the sequencing data may take the form of an electropherogram, as is known in the art, from which the DNA sequence is determined.

Where labeled primers are employed, the nature of the labeled primers will, in part, determine whether the families of labeled primer extension products may be separated in the same electrophoretic medium, e.g. in a single lane of slab gel or in the same capillary, or in different electrophoretic media, e.g. in different lanes of a slab gel or in different capillaries. Where the same labeled primer generating the same

detectable single is employed to generate the primer extension products in each of the different families, the families of primer extension products will be electrophoretically separated in different electrophoretic media, so that the families of primers extension products corresponding to each base in the nucleic acid can be distinguished.

Where different labeled primers are used for generating each family of primer extension products, the families of products may be grouped together and electrophoretically separated in the same electrophoretic medium. In this preferred method, the families of primer extension products may be combined or pooled together at any convenient point following the primer extension product generation step. Thus, the primer extension products can be pooled either prior to contact with the solid phase, while bound to the solid phase or after separation from the solid phase but prior to electrophoretic separation.

Kits for practicing the subject sequencing methods are also provided. At a minimum such kits will comprise capturable chain terminators, e.g. biotinylated-ddATP; -ddTTP; -ddCTP and -ddGTP. For embodiments in which the primer extension products are labeled, the kits will further comprise a means for generating labeled primer extension products, such as labeled deoxynucleotides, or preferably labeled primers, where the labeled primers are preferably Energy Transfer labeled primers which absorb at the same wavelength and provide distinguishable fluorescent signals. Conveniently, the kits may further comprise one or more additional reagents useful in enzymatic sequencing, such as vector, polymerase, deoxynucleotides, buffers, and the like. The kits may further comprise a plurality of containers, wherein each contain may comprise one or more of the necessary reagents, such as labeled primer, unlabeled primer or degenerate primer, dNTPs, dNTPs containing a fraction of fluorescent dNTPs, capturable ddNTP, polymerase and the like. The kits may also further comprise solid phase comprising a moiety capable of binding with the capturable ddNTP, such as streptavidin coated magnetic beads and the like.

1.4.7 Production of the DNA Fragments

In another embodiment, the DNA fragments are preferably prepared according to either the enzymatic or chemical degradation sequencing techniques previously described, but the fragments are not tagged with radioactive tracers. These standard procedures produce, from each section of DNA to be sequenced, four separate collections of DNA fragments, each set containing fragments terminating at only one of the four bases. These four samples, suitably identified, are provided as a few microliters of liquid solution.

1.4.7.1 Sample Preparation and Introduction

To obtain intact molecular ions from large molecules, such as DNA fragments, by UV laser desorption mass spectrometry, the samples should be dispersed in a solid matrix that strongly absorbs light at the laser wavelength. Suitable matrices for this purpose include cinnamic acid derivatives such as (4-hydroxy, 3-methoxy) cinnamic acid (ferulic acid), (3,4-dihydroxy) cinnamic acid (caffeic acid) and (3,5-dimethoxy, 4-hydroxy) cinnamic acid (sinapinic acid). These materials may be dissolved in a suitable solvent such as 3:2 mixture of 0.1% aqueous trifluoroacetic acid and acetonitrile at concentrations which are near saturation at room temperature.

One technique for introducing samples into the vacuum of the mass spectrometer is to deposit each sample and matrix as a liquid solution at specific spots on a disk or other media having a planar surface. To prepare a sample for deposit, approximately 1 microliter of the sample solution is mixed with 5-10 microliters of the matrix solution. An aliquot of this mixed solution for each DNA sample is placed on the disk at a specific location or spot, and the volatile solvents are removed by room temperature evaporation. When the solution containing the samples and thousand-fold or more excess of matrix is dried on the disk, the result should be a solid solution of samples each in the matrix at a specific site on the disk.

Each molecule of the sample should be fully encased in matrix molecules and isolated from other sample molecules. Aggregation of sample molecules should not occur. The matrix need not be volatile, but it must be rapidly vaporized following absorption of photons. This can occur as the result of photochemical conversion to more volatile substances. In addition, the matrix must transfer ionization to the sample. To form protonated positive molecular ions from the sample, the proton affinity of the matrix must be less than that of the basic sites on the molecule, and to

form deprotonated negative ions, the gas phase acidity of the matrix must be less than that of acidic sites on the sample molecule. Although it is necessary for the matrix to strongly absorb photons at the laser wavelength, it is preferable that the sample does not absorb laser photons to avoid radiation damage and fragmentation of the sample. Therefore, matrices which have absorption bands at longer wavelengths are preferred, such as at 355 nm, since DNA fragment molecules do not absorb at the longer wavelengths.

Depicted herein is a suitable automated DNA sample preparation and loading technique. In this approach, a commercially available autosampler is used to add matrix solution from container to the separated DNA samples. A large number of DNA fragment samples, for example 120 samples, may be loaded into a sample tray. The matrix solution may be added automatically to each sample using procedures available on such an autosampler, and the samples may then be spotted sequentially as sample spots on an appropriate surface, such as the planar surface of the disk rotated by stepper motor. Sample spot identification is entered into the data storage and computing system which controls both the autosampler and the mass spectrometer. The location of each spot relative to a reference mark is thus recorded in the computer. Sample preparation and loading onto the solid surface is done off-line from the mass spectrometer, and multiple stations may be employed for each mass spectrometer if the time required for sample preparation is longer than the measurement time.

Once the samples in suitable matrix are deposited on the disk, the disk may be inserted into the ion source of a mass spectrometer through the vacuum lock. Any gas introduced in this procedure must be removed prior to measuring the mass spectrum. Loading and pump down of the spectrometer typically requires two to three minutes, and the total time for measurement of each sample to obtain a spectrum is typically one minute or less. Thus 50 or more complete DNA spectrum may be determined per hour according to the present invention. Even if the samples were manually loaded, less than one hour would be required to obtain sequence data on a particular segment of DNA, which might be from 400 to 600 bases in length. Even this latter technique is much faster than the conventional DNA sequencing techniques, and compares favorably with the newer automated sequencers using fluorescence labeling. The technique of the present invention does not, however, require the full- time attention

of a dedicated, trained operator to prepare and load the samples, and preferably is automated to produce 50 or more spectrum per hour.

Greater detail of the preferred technique for DNA sequencing is depicted herein. Under the control of the computer, the disk may be rotated by another stepper motor relative to the reference mark to sequentially bring any selected sample to the position for measurement. If the disk contains 120 samples, operator intervention is only required approximately once every two hours to insert a new sample disk, and less than five minutes of each two hour period is required for loading and pumpdown. With this approach, a single operator can service several spectrometers. The particular disk geometry shown for the automated system is chosen for illustrative purposes only. Other geometries, employing for example linear translation of the planar surface, could also be used.

1.4.7.2 The Mass Spectrometer

The present invention preferably utilizes a laser desorption time of flight (TOF) mass spectrometer. The disk has a planar face containing a plurality of sample spots, each being approximately equal to the laser beam diameter. The disk is maintained at a voltage V_1 and may be manually inserted and removed from the spectrometer. Ions are formed by sequentially radiating each spot on the disk with a laser beam from source.

The ions extracted from the face of the disk are attracted and pass through the grid covered holes in the metal plates. The plates are at voltages V_2 and V_3 . Preferably V_3 is at ground, and V_1 and V_2 are varied to set the accelerating electrical potential, which typically is in the range of 15,000-50,000 volts. A suitable voltage $V_1 - V_2$ is 5000 volts and a suitable range of voltages $V_2 - V_3$ is 10,000 to 45,000 volts.

The low mass ions are almost entirely prevented from reaching the detector by the deflection plates. The ions travel as a beam between the deflection plates which suitably are spaced 1 cm. apart and are 3-10 cm long. The first plate is at ground and a second plate receives square wave pulses, for example, at 700 volts with a pulse width in the order of 1 microsecond after the laser strikes the tip. Such pulses suppress the unwanted low mass ions, for example, those under 1,000 Daltons, by deflecting them, so that the low weight ions do not reach the detector, while the higher weight ions pass between the plates after the pulse is off, so they are not deflected, and are detected by detector.

An ion detector is positioned at the end of the spectrometer tube and has its front face maintained at voltage V_d . The gain of the ion detector is set by V_d which typically is in the range of -1500 to -2500 volts. The detector is a chevron-type tandem microchannel plate array with a front plate at about -2000 volts. The spectrometer tube is straight and provides a linear flight path, for example, 1/2 to 4 meters in length, and preferably about two meters in length. The ions are accelerated in two stages and the total acceleration is in the range of about 15,000-50,000 volts, positive or negative. The spectrometer is held under high vacuum, typically 10 uPa, which may be obtained, for example, after 2 minutes of introduction of the samples.

The face of the disk is struck with a laser beam to form the ions. Preferably the laser beam is from a solid laser. A suitable laser is an HY-400 Nd-YAG laser (available from Lumonics Inc., Kanata (Ottawa), Ontario, Canada), with a 2nd, 3rd and 4th harmonic generation/selection option. The laser is tuned and operated to produce maximum temporal and energy stability. Typically, the laser is operated with an output pulse width of 10 ns and an energy of 15 mJ of UV per pulse. To improve the spatial homogeneity of the beam, the amplifier rod is removed from the laser.

The output of the laser is attenuated with a 935-5 variable attenuator (available from Newport Corp., Fountain Valley, Calif.), and focused onto the sample on the face, using a 12-in. focal length fused-silica lens. The incident angle of the laser beam, with respect to the normal of the disk's sample surface, is 70°. The spot illuminated on the disk is not circular, but a stripe of approximate dimensions 100×300 μm or larger. The start time for the data system (i.e., the time the laser actually fired) is determined using a beam splitter and a P5-01 fast pyroelectric detector (available from Molelectron Detector Inc., Campbell, Calif.). The laser is operated in the Q switched mode, internally triggering at 5 Hz, using the Pockels cell Q-switch to divide that frequency to a 2.5 Hz output.

The data system for recording the mass spectra produced is a combination of a TR8828D transient recorder and a 6010 CAMAC crate controller (both manufactured by Lecroy, Chestnut Ridge, N.Y.). The transient recorder has a selectable time resolution of 5-20 ns. Spectra may be accumulated for up to 256 laser shots in 131,000 channels, with the capability of running at up to 3 Hz, or with fewer channels up to 10 Hz. The data is read from the CAMAC crate using a Proteus IBM AT compatible computer. During the operation of the spectrometer, the spectra (shot-to-

shot) may be readily observed on a 2465A 350 MHz oscilloscope (available from Tektronix, Inc., Beaverton, Oreg.). A suitable autosampler for mixing the matrix solution and each of the separated DNA samples and for depositing the mixture on a solid planar surface is the Model 738 Autosampler (available from Alcott Co., Norcross, Ga.).

This linear TOF system may be switched from positive to negative ions easily, and both modes may be used to look at a single sample. The sample preparation was optimized for the production of homogeneous samples in order to produce similar signals from each DNA sample spot.

1.4.7.3 Data Analysis and Determination of Sequence

The raw data obtained from the laser desorption mass spectrometer 30 consists of ion current as a function of time after the laser pulse strikes the target containing the sample and matrix. This time delay corresponds to the "time-of-flight" required for an ion to travel from the point of formation in the ion source to the detector, and is proportional to the mass-to-charge ratio of the ion. By reference to results obtained for materials whose molecular weights are known, this time scale can be converted to mass with a precision of 0.01% or better.

In a graph of intensity v. time-of-flight of the pseudomolecular ion region of a TOF mass spectrum of Not I Linker (DNA) in which the matrix is ferulic acid and the wavelength is 355 nm, four consecutive spectra can be obtained using the present invention by the successive measurement of the four collections of DNA fragments obtained from fragmentation of each sample of DNA. Each of these spectra will correspond to the set of fragments ending in a particular base or bases G, G and A, C and T, or C. To determine the order of the peaks in the four spectra, a simple computer algorithm may be utilized.

It should be noted that the data obtained from the mass spectra contains significantly more useful information than the corresponding traces from electrophoresis. Not only can the mass order of the peaks be determined with good accuracy and precision, but also the absolute mass differences between adjacent peaks, both in individual spectra and between spectra, can be determined with high accuracy and precision. This information may be used to detect and correct sequence errors which might otherwise go undetected. For example, a common source of error which often occurs in conventional sequencing results from variations in the amounts of the individual

fragments present in a mixture due to variations in the cleavage chemistry. Because of this variation it is possible for a small peak to go undetected using conventional sequencing techniques. With the present invention, such errors can be immediately detected by noting that the mass differences between detected peaks do not match the apparent sequence. In many cases, the error can be quickly corrected by calculating the apparent mass of the missing base from the observed mass differences across the gap. As a result, the present invention provides sequence data not only much faster than conventional techniques, but also data which is more accurate and reliable. This correction technique will reduce the number of extra runs which are required to establish the validity of the result.

1.4.8 The Amplification Of A DNA Stretch Using The Pcr Procedure With The Knowledge Of Only One Primer

In another embodiment, the present invention enables the amplification of a DNA stretch using the PCR procedure with the knowledge of only one primer. Using this basic method, the present invention describes a procedure by which a very long DNA of the order of millions of nucleotides can be sequenced contiguously, without the need for fragmenting and sub-cloning the DNA. In this method, the general PCR technique is used, but the knowledge of only one primer is sufficient, and the knowledge of the other primer is derived from the statistics of the distributions of oligonucleotide sequences of specified lengths.

1.4.8.1 Method of Sequencing without the Need for Fragmenting or Subcloning

The objects and advantages of the present invention are also achieved by a method comprising:

- a) synthesizing a partly fixed primer, with 4, 5, 6 nucleotide, or longer sequence characters fixed within it. The fixed sequence can be any sequence, with some preferred sequences such as those containing many G-C pairs that increases binding affinity. The fixed position within the primer can be anywhere, with some preferred positions;
- b) taking a very long genomic DNA, either uncloned or a cloned large insert such as the YAC or cosmid in which a short sequence of about 20 characters somewhere within the DNA is known;
- c) synthesizing a primer from the sequence known from the DNA in step b;
- d) radiolabeling the primer in step c;
- e) annealing the primers (from step a, and step d or step g as appropriate) to the DNA in step b, and amplifying the DNA between the attached primers;
- f) performing DNA sequencing of the amplified DNA by the chemical degradation method of Maxam and Gilbert, or carrying out DNA sequencing by the Sanger method, or by modified PCR-sequencing method;
- g) after obtaining the DNA sequence from step f, selecting an appropriate first primer towards the 3' end of the sequence, synthesizing it, and radiolabeling it;

h) repeating the steps e through g with the two primers (the same partly fixed unknown primer as the second primer and the newly synthesized primer from step g as the first primer);

i) if the sequence obtained in step f is too short to be of value, using another partly fixed primer with a different fixed sequence and the same first primer to obtain a longer DNA sequence.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned hereunder are incorporated herein by reference. Unless mentioned otherwise, the techniques employed herein are standard methodologies well known to one of ordinary skill in the art.

The partly fixed primer used to perform DNA amplification and sequencing are, of course, not limited to those described under the examples. Further modification in the method may be made by varying the length, content and position of the fixed sequence and the length of the random sequence. Additional obvious modifications include using different DNA polymerases and altering the reaction conditions of DNA amplification and DNA sequencing. Furthermore, the basic technique can be used for sequencing RNA using appropriate enzymes.

Instead of preparing the first primer completely, it can also be prepared as follows. Two or three shorter oligonucleotides that would comprise the complete primer could be ligated, by joining end-to-end after annealing to the template DNA, as described under another patent (Helmut Blocker, U.S. Pat. No. 5,114,839, 435/6, 5/1992) or as described in the publication (L. E. Kotler, et al., Proceedings of the National Academy of Science, USA, 90:4241-4245 (1993)). Alternatively, it can be synthesized using the single-stranded DNA binding protein, the subject of another invention (J. Kieleczawa, et al., Science, 258:1787-1791 (1992)). One of such procedures, or an improved version thereof, can be used to make the first primer in the present invention. All in all, the first primer need not be synthesized at every PCR reaction while contiguously sequencing a long DNA, and can be directly constructed from an oligonucleotide bank. Based on the present invention, the second primer also can be chosen from a set of only a few pre-prepared primers. This enables the direct automation of sequencing the whole long DNA by incorporating the primer elements into the series of sequential PCR reactions.

1.4.8.2 Advantages of Method

An advantage of the present invention is that from a known sequence in a very long DNA, sequencing can be performed in both directions on the DNA. Two first primers can be prepared, one on each strand, running in the opposite directions, and the sequence can be extended on both directions until the two very ends of the long DNA are reached by the present invention, using a small set of pre-prepared partly fixed second primers.

One of the major advantages of the present invention is that it is highly amenable to various kinds of automation. Instead of radiolabeling the first known primer, it can be fluorescently labeled, and with this the DNA sequencing can be performed in an automated procedure on machines such as that marketed by the Applied Biosystems ("373 DNA Sequencer: Automated sequencing, sizing, and quantitation", a pamphlet from the Applied Biosystems, A Division of Perkin-Elmer Corporation (1994)). In the present invention there is no need to newly synthesize any primers to sequence a very long DNA. Thus, with the pre-prepared set of partly fixed second primers, an oligonucleotide bank for the synthesis of the first primer, and a large supply of the template genomic DNA (or any long DNA), the sequencing of the whole long DNA can be automated using robots almost without any human intervention, except for changing the sequencing gels.

1.4.8.3 Applications of Method

The following processes can be computer controlled: 1) the selection of the appropriate sequence for constructing the first primer close to the 3' end of the newly worked out sequence, 2) determining whether the sequence obtained is too short and selection of a different partly fixed second primer, 3) assembling the contiguous DNA sequences from the various lanes and various gels and appending to a database, and other such processes. Thus the present invention enables the construction of a fully automated contiguous DNA sequencing system. Any such automations are obvious modifications to the present invention.

The present invention is not limited to only unknown genomic DNA, and can be used to sequence any DNA under any situations. DNAs or RNAs of many different origins (e.g. viral, cDNA, mRNA) can be sequenced not only limited to research or information gathering purposes, but also to other purposes such as disease diagnosis and treatment, DNA testing, and forensic applications.

It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

It should be noted that any kit or process used for research, diagnostic, forensic, treatment, production or other purposes that uses the present invention is covered under these claims. Furthermore, the various sequences of the partly fixed second primers that can be used in the present invention are covered under this patent. Thus, any kit or process that uses this method and/or the DNA strands with the sequences that would comprise the partly fixed second primers will also be covered under this.

In addition to contiguous DNA sequencing, the present invention will cover the amplification of the DNA strands that are bounded between the known primer and the partly fixed second primer (either from claim 1 or from claim 2). The DNA amplification can also be performed for long DNA strands using the long PCR amplification protocols.

1.4.9 Polynucleotide Sequencing With Random Surface Immobilization And Light Microscopic Detection Of Affinity Labels Coupled To Microscopic Beads

A DNA sample is prepared by shearing or digestion at a first sequence with a first restriction enzyme producing a 3' overhang terminus, to some appropriate, known size distribution, and labeled with a digoxigenin bearing nucleotide by the action of terminal deoxynucleotidyl transferase.

After such digoxigenin labeling, said DNA sample is then subjected to random internal cleavage, for example by shearing so as to produce a population of molecules with an average length half that produced in the previous sizing step, or digestion with a second restriction enzyme recognizing a distinct, second recognition sequence. Sample molecules of said sample are then bound at some convenient surface density to a transparent surface modified with a monolayer or a sub-monolayer density of anti-digoxigenin antibody. Said sample molecules, which will thus be bound to said transparent surface by the 3' termini of one strand, are then subjected to treatment by a 3' to 5' exonuclease, which will only act at the 3' terminus which does not bear the digoxigenin moiety due to the hindrance of this latter 3' terminus by its interaction with the surface, preferably not to completion of digestion of susceptible strands. Thus primed DNA sample template molecules bound to a transparent surface in an end-wise manner are prepared.

Using a single nucleotide labeling affinity moiety in a manner similar to the example provided for one-bit binary labeling systems, utilizing for example each of the four nucleotides derivatized to effect communication of said nucleotides with a biotin moiety via a chemically cleavable linker, such as those described by S.W. Ruby et al.³⁴ polymerization directed by the template provided by each involved DNA sample template molecule is effected with an appropriate DNA polymerase lacking a 3' to 5' exonuclease activity, such as Sequenase 2.0,³⁵ with only one nucleotide type present during each polymerization step sub-cycle, at sufficiently low concentration to effect equilibrium controlled stepping. Polymerization reagents are then washed away, and may favorably be recycled after quantitation and readjustment of respective labeled nucleotide content.

After each such polymerization sub-cycle step, which will add a biotin labeled nucleotide to only a fraction of those sample template molecules having only the base complementary to the nucleotide of said sub-cycle located immediately 5' to the base

opposite the 3' terminal base of the strand priming this nucleotide addition, biotin bearing molecules may be labeled with microscopic streptavidin coated beads. Unbound beads are then washed away. Bead labeled molecules may then be observed by a video microscope, and the position of said bead labeled molecules within a sample may be recorded by image analysis of digital images thus obtained, in a manner similar to that used by Finzi and Gelles. Dithiothreitol or other reagents capable of cleaving said linker holding said biotin in communication with said nucleotide incorporated during the previous polymerization sub-cycle are then used to treat sample molecules to cleave said linkers and thus release said biotin labeling moieties and the beads which have bound to them. A wash step is then performed to remove said beads. The extent of bead removal may be checked with another video microscopy detection step if needed; and further cleavage treatment may be performed if decoupling was not adequate. The same subcycle (comprising polymerization, bead association, video microscopic examination, bead and label cleavage and removal by washing, and optionally a bead removal confirmation video microscopic examination step) is then repeated in succession for each of the three remaining nucleotide types, to complete a full base sequencing cycle (which as noted may yield information about more than one base location for some template molecules according to the sequence composition and the order of sub-cycles, and no information for other sample template molecules). Multiple said base sequence cycles are repeated until enough data have been accumulated relative to the total complexity of the initial DNA sample. Recorded data are then used to reconstruct sequence information for a segment of each sample template molecule, and segment sequence data are then aligned by appropriate computational algorithms.

Note that this embodiment avails only existing and generally available materials and devices, relies on relatively simple manipulations which are known to be highly reproducible according to their general use in the relevant fields, but due to the novel process of the present invention may yield genome sequence information far more rapidly and inexpensively than highly complex robotic instruments with sequencing methods utilizing electrophoretic separation.

Note that microscopic detection may be performed with a computer controlled steppable sample stage to effect the automated examination of large surface areas and hence very large numbers of sample molecules.

Alternatively, the transparent substrate providing the surface for immobilization may be that of a spooled film, which may be advanced at an appropriate rate before the objective of said video microscope of the present embodiment. Further, with such a spooled sample arrangement, said film may be circular, and continuously advanced through multiple video microscope apparatus and wells effecting polymerization sub-cycles, all in appropriate order such that benefit of full pipelining of each step may be enjoyed. The construction of such instrumentation and rudimentary robotic actuation systems will be straightforward to those skilled in the relevant engineering arts.

Surface immobilization with single photon detection of plural fluorescent labels coupled to photodetachable 3'-hydroxyl protecting groups. Sequence determination may additionally be effected by the random immobilization at some appropriate density of appropriately prepared and primed sample molecules on the surface of a transparent film, and stepwise polymerization with some appropriate polymerase, of all four nucleotides, all of which are protected at the 3'-hydroxyl with a photolabile (and hence photoremovable) protecting group in communication with labeling moieties which distinctly correspond to each nucleoside base type of the respective nucleotide. Label incorporation is detected, for example by the scanned beam light microscopic methods of the present invention, or with highly sensitive CCDs, and assigned to the spatial region occupied by a particular molecule. Said film is translated appropriately such that the full complexity of the sample may be examined after each polymerization cycle.

Data are recorded electronically and according to the molecule for which they are obtained. Illumination of the sample with an appropriate frequency and intensity of light to effect 3'-hydroxy deprotection and hence also labeling moiety removal is performed, and a wash step is performed to remove freed label. Such polymerization, detection and deprotection cycles are repeated until the sample is sufficiently well characterized.

1.4.9.1 Random And Non-Random Immobilization To Optical Detection Array Devices With Optical Labels

1.4.9.1.1 Detection And Classification Of Pathogens In Clinical Samples

Methods of the present invention may be combined with the immobilization of highly diverse libraries of binding specificities with either encoding labels or

phenogenocouples, which may therefore be characterized dynamically and related to any detected binding of particles of interest from a sample. Clinical samples are interacted with said libraries. All retained material is then interacted with some general label such as a polynucleotide binding dye (e.g. ethidium bromide, DAPI) or some chromophorigenic or photoemissive or labeled competitive inhibitor analog reagent detecting some metabolically fundamental reaction such as ATP hydrolysis, or the presence enzymes catalyzing said metabolically fundamental reaction. Pathogens containing polynucleotides or capable of said metabolically fundamental reaction may thus be detected.

The essential features of such a system are massively parallel screening for affinity interactions, generalized labeling methodology, and automated sample characterization. Because pathogen culturing is not required, and many types of highly specific information may be obtained in one assay procedure, without any previous knowledge of the state of the organism from which said clinical sample was obtained, this represents the basis for extremely powerful diagnostic methods.

Note that various implementations may distribute binding specificities of known composition in a spatially controlled manner, and thus rely on spatial information to encode specificity type and hence, if known, composition of each specificity type. Note also that said libraries may comprise known mimetics or small molecules of known binding specificity.

The profile of any sample type from an individual organism according to such an assay may be monitored over time, and a profile is preferably obtained for a state of presumed health for comparison to samples correlated to states of disease, deficiency or degeneration or other states of ill health (i.e. longitudinal tracking of individuals stratified by sample type). Samples of similar type may also be compared across populations and subpopulations, and the profile of these samples also correlated with state of health of the respective individuals (cross- sectional comparison).

For additional selectivity of detection, such a sample characterized as above may be further characterized according to the immunocharacterization method below.

1.4.9.1.2 Automated Immunocharacterization And Cyber-Immune Detection

Such a system resembles that used for the detection and characterization of clinical samples, except that said highly libraries of binding specificities comprises a large number of immunoglobulin specificities. Libraries comprising immunoglobulin specificities may include such specificities in the form of immunoglobulins expressed on bacteriophages, viruses, or in the form of the phenogenocouples of the present invention.

Banks comprising all of the specificities of a library may be maintained as monoclonal, and upon detection of a pathogen in association with one or more binding specificity contained in some library, and the identification and/or characterization of said one or more binding specificity, an alignment of the respective said monoclonal, from one of said banks, may be provided to the organism. Such analysis and provision of one or more monoclonal be automated and controlled by algorithms.

Similar rapidity and broad characterization advantages are attained as with the preceding method for the characterization of clinical sample.

1.4.9.1.3 Massively Parallel Enzymological Assays:

In a manner similar to the preceding embodiments, several enzymes contained within some sample may be analyzed according to their binding probability, binding duration or dissociation rate and conformational or phosphorylation or other status. Such assays may favorably be performed by the methods of the present invention, with immobilized libraries which may include competitive inhibitors, and with pre- or post-binding labeling of sample enzymes by encoded label antibodies, to permit classification of sample enzyme type on a molecule by molecule basis, which classification data may be combined with the data obtained in this assay.

1.4.9.2 Hybridization Based Detection Of Polynucleotide Sequences.

Various methods have been developed to test for the presence of short polynucleotide sequences and combinations of such sequences (according to stringency) in polynucleotide samples by hybridizing oligonucleotides or polynucleotides of known sequence to said polynucleotide samples. Such methods are sometimes termed 'gene-probe' methods and often involve the use of immobilized, ordered arrays of oligonucleotides of known composition.

Said ordered arrays have been formed on the surfaces of integrated electronic devices. It has been shown that, provided stringency can be made sufficiently high to prevent binding with even one base mismatch, such methods may be used to obtain sequence information about a sufficiently small sample.

The methods of the present invention provide a more rapid and convenient method for testing for the binding of known oligonucleotides to a complex polynucleotide sample, owing largely to the higher degree of parallelism which may be accomplished with single molecule methods. Here, each oligonucleotide, of known sequence, to be used as a specific gene probe, is synthesized with some perceptible encoded label, as described above, where the codes assigned to the sequence of said each oligonucleotide are known (due to the synthetic scheme by which they are produced and concurrently labeled). These are then hybridized to sample polynucleotide molecules, which either have previously been or will subsequently be immobilized, or may otherwise be separated from probe oligonucleotides, and the presence or absence of said each oligonucleotide in the sample polynucleotide containing fraction, which is a direct result of the success or failure of said each oligonucleotide to bind said sample polynucleotide molecules, will be readily ascertained through the detection and discrimination of the perceptible encoding labels corresponding to said each oligonucleotide. Contrary to the conventional gene-probe methodology, known probing molecules are generally unbound in this variation of the method as may be used with the present invention.

If the complexity of the polynucleotide sample is not too large, and the population made up of said oligonucleotides is sufficiently large and complex, preferably exhaustively enumerating all possible oligonucleotides of the respective and sufficiently long length, and provided hybridization may be sufficiently stringent, which stringency is affected by a large number of known factors but also has sequence dependent components, information about the binding of said each oligonucleotide, which may be related to the respective known sequence and by Watson-Crick pairing rules to the respective sample polynucleotide sequence segment (or by identity with the strand complementary to the strand to which said each oligonucleotide has bound) may thus be obtained. As with other methods, alignment of such data may yield information about the sequence of the sample. The methods of the present invention further provide for the quantitation of such oligonucleotide

hybridization by way of counting the number of times a particular perceptible encoded label is retained by a said polynucleotide sample, which may be availed both in the monitoring and correcting of errors and in the modulation of binding (hybridization) conditions.

Alternatively, probing may be accomplished by oligomeric sequences immobilized in some known configuration, for example by spatially patterned methods such as those of S.P.A. Fodor et al.³⁷ or by the lattices produced hierarchically by the method of N.C. Seeman noted above but comprising an ordered array (the order of which is predetermined by the incorporation or association of single stranded oligonucleotides or other single stranded termini of known sequence into or with modular components used to build up said lattices) of short single stranded regions of known sequence and preferably one free terminus (so as not to hinder conformational changes required for hybridization), but detected by the methods of the present invention, where sample polynucleotides are labeled with some appropriate discernible label, such as the dye YOYO-I, to facilitate the detection of their presence in association with each of said oligomeric sequences.

A yet further variation for effecting the spatially predetermined distribution of, for example and exhaustively enumerated population of single stranded oligonucleotides, may be effected by the use of the methods of N.C. Seeman to produce a uniform two dimensional lattice with a repeating pattern of short single stranded sequences with photo protected termini, for example all of the 256 possible 4-mers. Such a lattice may have a periodicity substantially smaller than the wavelength of visible light. Said short single stranded sequences may comprise some synthetic backbone so as to be resistant to enzymatic cleavage, which backbone preferably also is non-ionic (for example, of alkyl or beta-cyanoethyl derivation, peptide-nucleic-acid composition, or methylphosphonate composition) so as to denature from a complementary sequence only at markedly elevated temperatures relative to ordinary oligonucleotides. Thus, a pattern of oligonucleotide complexity may be distributed in a predetermined manner below the resolution of light directed patterning.

Light patterning techniques may then be availed to spatially direct the photodeprotection of said short single stranded sequences at lower resolution. Such light directed syntheses are preferably terminated with some comonomer which will

prevent exonucleolytic degradation of said short single stranded sequences, or all of said short single stranded sequences are of a polarity opposite to that specified by the exonuclease to be subsequently used. By this combination of methods, patterning resolution is not limited by the properties of light, but may avail of the convenience of light directed patterning at lower resolutions. After a known distribution of all possible single stranded sequences of sufficient complexity has thus been produced, a denatured, labeled polynucleotide sample produced by extensive nick translation, with fluorescent labeled nucleotides, of a naturally occurring polynucleotide sample is hybridized to said lattice. Hybridized molecules are treated mildly with a single strand specific nuclease, followed by an exonuclease, to degrade or by the same process to free those regions which are not bound to the probing said short single stranded sequences. Label incorporated into the nick translation products of said polynucleotide sample is then detected and spatially mapped by the methods of the present invention, and binding is thus scored according to the known probing said short single stranded sequences. This method thus avails the molecular parallelism made possible by the molecular recognition, high density and high resolution detection methods availed with the present invention.

Note, finally, that higher density patterning than attainable by conventional light patterning methods may also be effected by scanning probe lithographic methods, such as the use of NFSOM lithography with photodeprotectable groups.

1.4.9.2.1 Methods For Repeatable Detection And Identification Of Single Molecules

Repeatable detection and identification of single molecules is achievable by microscopic labeling with some readily identifiable, e.g. combinatorially or permutationally diverse and readily examined particle or molecule or group of molecules and detection of the thus marked identity of individual free molecules in solution, with removal of excess nucleotides (e.g. by filtration); and, scanning of a liquid sample volume where sample molecules and sample conditions are matched to ensure manageably slow free diffusion of sample molecules permitting tracking of the motions of free individual molecules in solution, as observed by T.T. Perkins et al. for reptation of DNA in solution, in which instance unreacted labeled monomers may be removed, for instance, according to their more rapid diffusion, possibly through a

filter, and detection may favorably comprise observation of reduced mobility of a labeling moiety after it has become attached to a sample molecule.)

According to the labeling methods employed, various detection methods may satisfy the requirements of signal detection with repeatable assignability to a particular unique sample template molecule.

Prominent among these detection methods are microscopy methods such as video microscopy including confocal fluorescence microscopy with or without enhancement, and with or without variations incorporated into the present invention near field scanning optical microscopy (NFSOM) and variations thereof; contact and non-contact varieties of scanning force microscopy (SFM; also termed atomic force microscopy (AFMI) and variations thereof; other scanning probe microscopies including scanning tunneling microscopy (STM), scanning tunneling spectroscopy (STS), and so-called field emission mode STM (which is more accurately described as microscopy by field emission from a scanned conductive probe, or scanning field emission microscopy, SFEM, because no tunneling actually occurs). Any enhancements of scanning probe microscopy, including multiple probe parallelism, may readily be availed in the practice of the present invention.

Additionally, optical detection methods employing optoelectronic array devices (OADs), such as spatial light modulators (SLMs), laser diode arrays (LDAs), light-emitting diode arrays, or charge coupled photo-diode arrays (conventionally termed CCDs), in combination with appropriately high sensitivity detection methods, may also be employed, particularly with samples immobilized such that the maximal proportion of pixel elements of said array will be involved with the detection of a signal from exactly one sample molecule. CCD and SLM array device are presently available at pixel densities of approximately 10^5 to 10^6 per cm^2 . LDAs of comparable density are currently under development. Device level constraints upon parallelism will thus be significant, but may be overcome by increasing the data obtained per molecule (i.e. processivity or sequence segment length.) Such devices may be employed remotely, i.e. in some arrangement where light passes through the sample under study and is detected by some apparatus involving said array devices, or in close or direct contact with said sample, as for instance, polynucleotides have been immobilized to integrated circuits for other applications. Appropriate arrangements of such devices for the appropriate detection scheme in which each device type is

appropriately used will be obvious to those skilled in the arts of optics and optoelectronics.

Note that for purposes of those variations of the present invention involving the immobilization of sample molecules, said immobilization may be conveniently effected in a random manner, relying upon some appropriate surface or volume density which yields a corresponding random surface or volume distribution, and appropriate detection methods to permit repeatable resolution of most sample molecules from each other. The length of the molecules in question will be an important factor in the determination of a desirable said density. Generally speaking, for random surface immobilization and without the use of measures to orient or order sample molecules, for molecules of length L (which may additionally account for any labeling bead diameter), and detection methods relying on spatial resolution R , maximum practical molecule number density will generally be the less than $1/(2L+R)^2$. This assumes the worst case configuration of two end immobilized molecules extending directly towards each other and both labeled near their respective termini. Similar calculations may be applied to three dimensional cases. Alternatively, one may consider $(2L+R)^2$ or $(2L+R)^3$ to be an average bin size, and determine via the Poisson distribution the optimal molecular number density corresponding to the largest number of bins being occupied by precisely one sample template molecule.

Alternatively, molecules may be labeled by a first label, for example with a particular fluorescent dye incorporated by nick translation, in a manner identifying a portion of the molecule near the site of polymerization, and proximity of said first label to the perceptibly distinct labeling moieties used for nucleotide incorporation detection and discrimination will permit the detection of unacceptable proximity of two distinct sample molecules. Such a method is consistent with the tracking methods described below for free sample molecules. In such a case, the data collected during the cycle in which said unacceptable proximity is observed for the affected molecules may be ignored, and lack of information from this cycle noted for the respective molecules. Conditions, such as solution viscosity, sample molecule diffusion rate, sample molecule concentration, sample dimensions, etc., may be optimized to reduce the occurrence of such unacceptable proximity, and oversampling methods described in other portions of the present disclosure may be applied to preclude this form of

error from degrading final data quality. These methods may be applied to either immobilized or unimmobilized sample molecules.

1.4.9.2.1.1 Microscopy Based Detection

Light microscopic visualization represents a particularly convenient and technically simple detection and unique molecule localization method. A visualization method of particular interest for purposes of the present invention in higher performance or more demanding applications is video enhanced confocal fluorescence microscopy (VECFM), preferably utilizing optics well matched to the refractive index of the reaction or detection medium.

As discussed above, various scanning probe microscopies may also be advantageously used within the present invention according to labeling agents and methods used. Most prominent among these are NFSOM and variations thereof, and both contact and non-contact SFM, and variations thereof.

Generally speaking, a microscopy based detection method must be sufficiently convenient, capable of use with a stepper translated or otherwise translatable sample, not destructive of the sample, and capable of detection of any labeling methodology to be used with it. Thus, it is quite likely that many microscopy methodologies not yet developed may readily be employed with the present invention. Further, microscopy and corresponding apparatus shall comprehend any miniaturized or microfabricated microscopy devices or other comparable integrated detection means.

1.4.9.2.1.2 HIGH SENSITIVITY AND SCANNED EXCITATION BEAM FLUORESCENCE CONFOCAL MICROSCOPY

A modification of VECFM which is particularly suited for SMD and SMV relies upon selective fluorescent excitation of an appropriate dye molecule label (or of molecules within a sample with appropriate fluorescent properties independent of labeling) in some sample by means of some tightly defined beam, with dimensions at or near the resolution limit of the apparatus, of an appropriate frequency, or of parametrically controllable frequency, where said beam is caused to scan in a controlled manner through the sample region within the visual field. This microscopy, including numerous variations, may be termed either scanned beam confocal microscopy or steered beam confocal microscopy (in either case, SBCM). Scanning

of said beam through the sample within the visual field may be accomplished by introducing said beam into the optical path of the VECFM via mobile mirrors which may effect said controlled scanning, or by first producing said beam with a pinhole which is itself scanned, before deflection towards the sample via said mirrors, which in the present case may be fixed in position, through the use of pinholes in a rotating disk arranged in one or more spiral arms to effect an approximately rastering illumination of the sample as said disk rotates, or by other means which will be obvious to those skilled in the design of optical instrumentation and microscopy. Said beam will excite fluorescence in any appropriately responsive molecules which occur in its path. An optical splitter may then redirect a fraction of the light transmitted from the sample through the objective lens, and direct it through a narrow bandwidth, high transmissiveness filter, which may be specific for a fixed or for a parametrically controllable variable frequency, to uniquely select the appropriate fluorescent emission frequency, to a highly sensitive photodetector, which may record either intensity as intensity information or as the number of photons detected per unit time, as a function of the region being subjected to fluorescence exiting illumination or being distinctly observed (see below). Thus a high resolution map of the fluorescence of the sample may be reconstructed, and further overlaid images obtained for the same sample and sample location by conventional VECFM means.

Alternatively, the entire sample of visual field may be subjected to illumination by an appropriate excitation frequency, and a pinhole scanned through the portion of the output of said optical splitter, such that light passing through said pinhole will reach said highly sensitive photodetector.

In yet a third, albeit technically more complex implementation, an SLM, may be used in place of said pinhole (in either configuration), and fluorescent excitatory illumination may be either broadly distributed or scanned.

In a fourth, albeit technically more complex implementation, sensitive photodetection may be accomplished with a highly sensitive CCD, and fluorescent excitatory illumination may be either broadly distributed or scanned. At present, CCD sensitivity approaching single photon detection is technically possible though is not practical for high volume applications.

In a fifth implementation, said scanned beam may originate from a laser diode array device or a light emitting diode array device, where only one of, or a contiguous

group of elements of, such an array is active at any particular time so as to produce a particular beam, and the group of active elements of said such an array is changed as a function of time to effect scanning of the sample by the coordinated activation and deactivation of the plural beams thus produced.

In all of the above implementations, spatial information is gained about any particular fluorescent emission, and this may then be combined with other visual information obtained via the same VECFM apparatus.

Note that for scanned beam methodologies, where beams are used for excitation or detection, even where said beams may have inhomogeneous but invariant distribution of internal flux density, known samples such as individual dye molecules may be imaged for calibration purposes and information useful for algorithmic enhancement may be collected. This information represents the characterization of the convolution of the beam and optics properties with the signal actually owing to the known sample, and thus localization of fluorescent sample features may be accomplished at better than optical resolution limitations. For example, a single, immobilized fluorescent molecule may be examined by such an apparatus, and the intensity as a function of beam position may be recorded for the full duration of its presence within the beam's path as said beam scans the sample, and the data thus obtained may then be used to determine the change in observed intensity as the sample molecule enters the extremity of the beam, traverses the beam and exits the beam. This information may then be subjected, for instance to averaging or other computations to determine the relationship between the location of the molecule within the beam and the intensity observed, and finally that information used to estimate the intensity which would be observed when such a calibration sample molecule is in the precise center of the beam. This information may then be used in image enhancement of unknown samples. Note, however, that localization to below optical resolution limitations is distinct from increasing the resolution capability for two nearby objects.

Scanning beam microscopies will be of particular advantage where it is desirable to use particular illumination frequencies to modify the sample. For purposes of the present invention, a beam of predetermined frequency, for instance delimited and scanned by means of a pinhole as described above, may be used to selectively modify a particular sample molecule. For example, a beam of

predetermined frequency may be used to effect the photobleaching of the labeling moiety on a particular sample molecule, to selectively remove a photocleavable protecting group on a particular sample molecule, to selectively remove a moiety joined to a sample molecule by a photocleavable linker, or selectively control any photochemical reactions in a highly localized but non-invasive manner.

Note that implementations permitting variations of illumination frequency and/or variations of the frequency or frequencies selected b,, * filters for detection purposes constitute microspectroscopy or microfluorimetry, and may be applied to any of the various light microscopies.

1.4.9.2.1.3 REPEATABILITY BY IMMOBILIZATION WITH DISCERNIBLE LOCATION

Surface Immobilization

A large number of methods presently exist to effect the immobilization of macromolecules and other molecules to various surfaces including the, surfaces of optically transparent materials. In general, such methods on the chemical modification of said surfaces such that they will be reactive with or have specific affinity for particular chemical functional groups placed on said macromolecules or molecules.

Applicable methods include those described by S.P.A. Fodor et al effect micropatterned surface immobilization and controlled synthesis polypeptides and polynucleotides, those described by M. Hegner et al.¹⁴ effect the end-wise immobilization of terminally thiol modified double helical DNA molecules to a gold coated surface, or those methods recently used by L. Finzi and J. Gelles¹⁵ to effect end-wise attachment of DNA molecules to an antibody coated glass surface. Many alternative methods will be obvious to those skilled in the relevant arts.

For purposes of genome sequencing applications of the present invention, DNA from a cosmid library which may have been prepared from total genomic material, from a cDNA library derived from a particular tissue type, from a cosmid library which may have been prepared for a single chromosome or group of chromosomes or particular chromosome segments, or directly purified genomic DNA or directly purified RNA from a particular cell type, etc., may be subjected to fragmentation. Physical methods such as shearing with a hypodermic apparatus may be suitable. Where the sample is in the form of duplex DNA, it may be treated with restriction enzymes, which preferably restrict either 6- or 4-base recognition sequences, so as to produce sample molecules of mean length of either 4 kilobases or 256 bases, respectively. Such lengths are sufficiently short to yield a high number density of sample molecules. Said sample molecules may then be appropriately derivatized, for example by fill-in reactions at 5' overhang cohesive termini produced by said restriction enzymes with nucleotides bearing an affinity label or an appropriately reactive chemical functional group.

1.4.9.2.1.4 MATRIX IMMOBILIZATION

There has been increasing interest and progress in the field of affinity chromatography which relies upon varying specific affinity interactions between molecules immobilized to a chromatographic matrix or polymeric matrix and the molecules contained in some sample. Of particular relevance are matrices with polynucleotides immobilized thereupon. An example which is widely known and used within the relevant fields is oligo-dT cellulose. Further, many chemistries and methods used to immobilize macromolecules to surfaces will be similarly applicable to immobilization to a polymeric matrix provided said matrix is chosen so as to have appropriate reactivities and not pose any difficulties associated with non-specific interactions. Most methods capable of effecting such matrix immobilization will be acceptable for purposes of the present invention. Note, however, that any matrix used in the present invention must admit the sufficiently rapid transport or diffusion of reagents, enzymes and buffers, as required by the particular embodiment.

1.4.9.2.1.5 FOCAL PLANE SCANNING

For detection and discrimination within a volume, whether for matrix immobilized samples or diffusion constrained free molecules in solution, especially where fluorescent labeling of one form or another has been employed, a sample may be examined by microscopy with reconstruction of three-dimensional spatial information by scanning the focal plane through the depth of the sample and collecting image data at appropriate intervals. Such methods of three-dimensional reconstruction are well known within the art of microscopy.

1.4.9.2.1.6 PLANE EXCITATORY ILLUMINATION

Alternatively, optical means such as moving slits or SLMs or laser diode arrays may be employed to selectively illuminate a particular region, preferably a single plane (of thickness similar to the wavelength of light employed or feature size of integrated device means employed), to examine a particular subset of sample template molecules and labels associated with them, providing spatial reconstructability of the data thus collected.

1.4.9.2.2 TWO BEAM METHODS INCLUDING PLANE ILLUMINATION

Volume distributed samples may also be examined with methods closely analogous to those recommended for three dimensional optical mass data storage, for instance, by Sadik Esener in U.S. Patent Number 5,325,324. Here, labels requiring excitation by photons of two distinct frequencies for photoemission may be employed. Alternatively, the related methods of illuminating an entire plane of a sample with one of said distinct frequencies may be availed as a mechanism for imaging with spatial reconstructability.

1.4.9.2.3 Immobilization Via Concatenation

For the various applications of the present invention involving the interaction of enzymes with extended linear macromolecules such as polynucleotides, when said extended linear molecules may be conveniently circularized by appropriate treatments (which will generally be obvious to those skilled in the relevant arts), immobilization of said extended linear molecules may be conveniently effected by their concatenation with second extended linear molecules which are likewise conveniently circularized by appropriate treatments (which will again generally be obvious to those skilled in the relevant arts) bearing chemical properties (i.e. functional groups such as thiols or affinity moieties such as biotin) favorable for convenient, specific immobilization to a surface, matrix or other solid support. For purposes of, for example, certain sequencing applications of the present invention, said second extended linear molecules are favorably bound (with methods which will generally be obvious to those skilled in the relevant arts) at a predetermined location along their length, to some protein, which may be an enzyme such as a polymerase, before immobilization. Said second extended linear molecules may have termini with reactive chemical functional groups which may be bound together by the addition of some appropriate reagent such as a chemical cross-linking agent, or with some affinity moiety such as an oligo- or polynucleotide which may be bound together by an appropriately complementary oligonucleotide or polynucleotide (with or without ligation thereof), or some appropriate multifunctional binding protein or receptor. Such an arrangement permits the following steps to be performed: said second extended linear molecule is bound to said enzyme; said protein is caused to bind to said first extended linear molecule (which may be circularized either in a prior or subsequent step); said second extended linear molecule to which said protein has been bound is caused to circularize by appropriate treatment; and if said first extended linear molecule is at this stage

linear, it is caused to circularize. Without any special measures, there is a fifty percent chance that such a process will result in concatenation of the first extended linear molecule with the second extended linear molecule. Numerous methods, such as size separation followed by retention by immobilization, may be used to purify the resulting desired concatenate. Where said second extended linear molecule was chosen to be relatively short, such an assemblage will provide for the retention of said first extended linear molecule, now in concatenated circular form, in proximity to said protein, with specific immobilization or convenient immobilizability. Thus, said protein and said first extended linear molecule now in concatenated circular form have a high effective concentration with respect to each other upon dissociation, and said protein and said first extended linear molecule now in concatenated circular form will not interact with the molecules of other such assemblages when said assemblages are at sufficiently low density or said second extended linear molecule now in concatenated circular form is particularly short (i.e. effectively shackles said first extended linear molecule now in concatenated circular form to said protein whether or not said first extended linear molecule now in concatenated circular form is bound by said protein.)

Such an immobilization scheme will be particularly desirable in, for example, sequencing applications of the present invention where a polymerase must perform a cycle, in which it binds, modifies and releases a sample molecule, at a high rate. A particular instance in which such desirability obtains is for samples to be analyzed with long sequence segments (e.g. hundreds or thousands of bases) where dissociation of the polymerase is necessary to permit either 3' hydroxy deprotection (e.g. removal of a photolabile protecting group) and or labeling moiety removal by appropriate means. Note that by immobilizing the enzyme, and hence the spatial location at which the labeling moiety first comes into physical communication with a sample molecule, the above stated limitation on sample molecule density may be overcome, with the new limit being that imposed by the detection method, thus increasing sample density and in some embodiments the parallelism that thence may readily be achieved with detection methods such as microscopy. It is therefore feasible, with such assemblages, to collect sequence data dynamically from each molecule at a rate approaching the limits imposed by the slower of: the characteristic nucleotide incorporation rate of the polymerase; or, the diffusion rate limit of nucleotide association with the nucleotide

binding site of the polymerase (divided by four) when nucleotides are at a sufficiently low concentration that their presence as labeled but free molecules in the detection field does not interfere with the detection (which may be time averaged according to the particular instrumentation used) of incorporated labeled nucleotides, which concentration will be dependent in part on the geometry of the liquid volume; or, the maximum rate of single label detection (but note that such a rate need not be low because detection rate will increase for multimeric labels, which may be employed). Such an immobilization method will favorably be employed for embodiments locating sample molecules on or near the surface of a CCD or SLM. Note that kinetic control of polymerization rate (and hence stepping rate, e.g. by adjusting nucleotide concentration) is also enhanced by the use of such a concatenation methodology.

1.4.9.3 IMMOBILIZATION WITH NON-RANDOM DISTRIBUTION

While the above methods are convenient precisely because they require only the simple optimization of sample molecule density, the resulting random distribution will less than fully utilize available substrate or matrix space and fewer than all sample molecules will be sufficiently well separated for unambiguous resolution of two adjacent sample molecules. Due to the inherent advantages provided by molecular parallelism, this will not in general be a significant constraint. For applications in which a high degree of instrumentation miniaturization is desired, however, a better effective density of usable sample molecules, distributed in either two or three dimensions, may be effected as needed by non-random immobilization methods.

One such random immobilization method may avail of the invention of N.C. Seeman, described in U.S. Patent Number 5,278,051, which provides a process for the construction of complex geometrical objects. These methods may be applied to the production of regular two- and three-dimensional molecular lattices from polynucleotide compositions. The process of this invention may be extended by the incorporation of appropriate affinity groups at predetermined locations within the objects, which for present purposes may favorably be small ligands such as biotin or digoxigenin, which may then be used as the target for a sample molecule which has been terminally labeled by a similar small ligand which has subsequently been bound by (an excess of) an appropriate multimeric receptor. Said multimeric receptor will

then recognize and bind the complementary small molecule ligand incorporated into the structure of said lattice, and thus effect sample molecule immobilization according to the non-random pattern predetermined by the precise structure of said lattice and the precise distribution of ligands thereupon. Note that because the objects provided by the invention of N. C. Seeman comprise polynucleotide structures, care must be taken in using such a sample substrate with the methods of the present invention to ensure that said objects will be stable to all treatments which are to be applied to sample molecules, including denaturation, exonucleolytic degradation, primer hybridization, exposure to active polymerases, etc. Generally, these constraints may be met by effecting topological closure of all strands such that no free polynucleotide terminus is carried on such a lattice, and no denaturation procedures will result in matrix dissociation; the methods of the invention of N.C. Seeman may be availed in a manner meeting these constraints.

Note that to ensure complete regularity of lattices constructed by such means, or any other molecular lattices which do not have complete internal rigidity, the extremities of these lattices may be bound to solid supports which are then positioned so as to apply tensile stresses to said molecular lattices which will enforce constraints limiting flexural internal degrees of freedom and enforcing substantial spatial regularity on sample molecule distributions.

Any other method which provides a regular array of binding sites to which sample molecules may selectively be associated will also suffice for the purpose of non-random immobilization of sample molecules in two- or three-dimensions for the present invention.

Note also that said appropriate affinity groups incorporated (directly or, by conjugation or other methods, indirectly) at appropriate sites in a lattice may be chosen so as to interact directly with polynucleotide sample molecules in a sequence dependent or independent manner. Sequence dependent affinity binding may be effected with oligonucleotides or analogs thereof capable of forming double-, triple- or quadruple helices with said sample polynucleotides, ribozymes, or sequence dependent binding proteins including but not limited to: transcriptional activators (e.g. TATA- Binding Protein), enhancers and repressors; integrases; restriction enzymes; replicator proteins (e.g. DnaA); DNA repair proteins; anti- polynucleotide antibodies, RNA processing complexes (e.g. snRNPs); and RNA binding proteins all under

conditions permitting desired selectivity, specificity or stringency but, where appropriate, preventing polynucleotide cleavage or degradation. Where sequence specific binding is desired, and hierarchically prepared lattices are used, the distribution of particular specificities may be controlled by the staged incorporation of said affinity groups at various hierarchical levels of the synthetic procedure. This will permit classification of sequence data according to the location of the sample template molecule from which it is obtained in the lattice (i.e. on the surface or within the matrix). Sequence independent binding of polynucleotides may be effected by the use of proteins such as RecA, histones, UI, etc.

1.4.9.3.1 Repeatable Identification Of Unimmobilized Molecules:

Single molecule tracking with controlled diffusion- For samples under continuous observation, e.g. continuously within a visual field of a video microscope, molecules may be perceptibly labeled, for example by perceptible microscopic beads or the incorporation of a first fluorescent label, and tracked by the use of image analysis algorithms. Said algorithms will recognize only the appropriate type of label and track the motions of the respective sample molecule as it slowly diffuses in solution, so as to permit the unambiguous direct correlation or assignment of the signal associated with the addition of a labeled nucleotide to said respective sample molecule. For these methods, nucleotide labeling does not necessitate the use of large beads or other complexes for detection. Instead, single or oligomeric fluorescent labeling moieties, or enzymatic label affinity conjugation are preferred, such that labels may be removed without greatly disturbing the trajectory of said respective sample molecules. Either the direct colocalization (to within the resolution of the imaging method) of nucleotide label with said first fluorescent label or reductions in the Brownian motion of said nucleotide label sufficiently near (e.g. closest to) said first fluorescent label may be exploited in the detection of nucleotide label incorporation.

Note that manipulation with a laser trap, as for instance described by T.T. Perkins et al. for reptation of DNA in solution, may be employed with such free molecules.

1.4.9.3.2 Unique Labeling Of Sample Molecules And Identification Methods

Various methods may be employed to uniquely label individual sample molecules. The complexity of such unique labels must be greater than the number of

sample molecules contained within a unitary sample preparation, such that any label is highly unlikely to occur more than once within said unitary sample preparation.

Labels may be visually discriminatable, or may be diverse affinity labels or combinations thereof. Labels of this type may conveniently be random combinations of some basis set of distinct labels, formed for example, by a random coupling or polymerization of such labeling moieties to a defined chemical site provided by chemical modification of sample molecules.

Visual labeling may be accomplished by the use of a sufficient number of distinguishable fluorescent dye molecules, or other visual labels, such that the presence or absence of association of any one of said distinguishable fluorescent dye molecules may comprise the state of a bit in a binary code. Such labeling is similar to the combinatorial encoding described by S. Brenner and R.A. Lerner, but differs in that: perceptible labels may be used for encoding; labels need not be genetic material or linear copolymers; where only unique identifiability is required, the label moiety employed for encoding may be synthesized separately and possibly randomly, and bound possibly randomly with sample molecules; the information contained by each labeling moiety need not depend on its precise spatial association with sample molecules, or its location within a sequence, only its sufficient proximity; and, because of such modes of independence between the encoding, which serves here only for purposes of unique labeling, difficulties which may arise for particular orthogonal polymerization chemistries of different copolymer types may be avoided either by separate synthesis. Alternatively, for biopolymers, and, possibly for specifically encoded libraries, the use of specific enzymes which may for example ligate polynucleotides or polypeptides, may be used to specifically control reactions and prevent polymerizations of one biopolymer from affecting a second, linked biopolymer. Note that moieties different from biologically occurring comonomers may be used as encoding: label moieties, via functionalization of appropriate biopolymer segment with such moieties, in synthetic manners which will be obvious to those skilled in the relevant arts, or may be used, similarly, as constitutes the random library thus encoded. This latter case is, for example accomplished with the use of multiple distinct short double stranded DNA molecules with appropriately complementary cohesive termini which each carry some particular affinity or photolabel type, and which may be ligated together in a manner stepped by the

addition of appropriate adaptor linkers, even in the presence of other biopolymers (such synthetic methods being further favorably facilitated by the use of solid phase synthetic methodologies). Depending on the sensitivity of the detection methods used, multimers of each single type of fluorescent dye moiety, or detectable multiplications of other photolabels, may be used to effect higher modulo coding of labels.

1.4.9.4 ENCODING BY SYNTHESIS WITH MULTIMACROMONOMERS

Note that the labeling methods of the present invention suggest a convenient solution to the problem recognized by Brenner and Lerner, as limiting the facility of their encoding system, i.e. the requirement of separate distinct comonomer (or co-oligomer) type addition steps for each polymer type. This prevents the use of highly random (but step- controlled) synthetic preparation of such encoded libraries, because the information encoded is realized by individual preparative synthetic steps, i.e. all of the information content of the encoding is conferred upon these compounds by the intervention or agency of a chemist (or automated systems) at each step. Such encoded libraries, of either the sequence encoded or modulo encoded types, including compounds comprising more than two polymer types, may be prepared with the following stepped random method in one container (with or without the favorable use of solid phase synthetic methodologies). Note that the term random here refers to the mixture of two or more multimacromonomers in each addition step, such that addition to all compounds under preparation will occur in a random manner within the reaction mixture, in a manner weighted according to the relative concentration of each such multimacromonomer. Such multimacromonomers may also be used in more directly controlled addition schemes with advantages which will be obvious to those skilled in the relevant arts.

Multimacromonomers comprising two or more monomer (or macromonomer) types (e.g. comprising an amino acid monomer and a trinucleotide oligomer, or an amino acid monomer, a trinucleotide oligomer and a fluorescent or affinity labeling moiety) may be prepared by joining some or all of said two or more monomer (or macromonomer) types by cleavable linkers such as those described in other sections of the present disclosure. Thus, each multimacromonomer may be added to compounds under synthesis by addition of one of the monomer or macromonomer types to the corresponding polymer or macropolymer types of said compounds under

synthesis by appropriate polymer synthesis chemistry, followed by addition of some or all of each of the remaining monomer or macromonomer types to the respective corresponding polymer or macropolymer types of said compounds under synthesis by appropriate polymer synthesis chemistry. Control over the details of such additions may be effected by control over, for example, removal of distinct protecting groups from distinct polymer or macropolymer types of said compounds under synthesis by appropriate polymer synthesis chemistry. Linkers or specific linker branches may be cleaved at appropriate steps or after synthesis has otherwise been completed. Thus, correspondence between the composition of each polymer or macropolymer type comprised within each molecule of the compound under synthesis (which final composition may vary widely from molecule to molecule of the compound under synthesis, but strictly observe the correspondence between composition of some or all of each of the polymers or macropolymers comprised within each molecule of the compound under synthesis) is provided by the communication of the distinct monomer or macromonomer types comprised within each multimacromonomer. The first bond formed between a first monomer or first macromonomer of a multimacromonomer and a molecule of the compound under synthesis will thus ensure that other monomer or macromonomer types of the multimacromonomer which will be added at the respective multimacromonomer addition stage will correspond to the identity of the first monomer or first macromonomer thus added. Thus correspondence of some or all of each of the polymer or macropolymer types of final compounds is enforced (by the communication effected by, for example, linkers) even where the composition of some or all of the polymer or macropolymer types is respectively random.

Preferably, such linkers (which may be multiply branched, each of such branches possibly comprising cleavable groups susceptible to distinct cleaving treatments) are held in communication with some or all of the two or more distinct monomer or macromonomer types (which are added to the compounds under synthesis with distinct and mutually non-interfering addition or polymerization, deprotection and/or activation chemistries, termed "orthogonal" chemistries in the respective art) by attachment to the protecting groups used to effect the stepping of additions of each such multimacromonomer. Said diverse affinity labels may be used in conjunction with multiple affinity separation paths and nucleotide label detection

that associates the detected said nucleotide label with the resolved location of the respective affinity labeled sample molecule, thus accomplishing the required assignment of detection and discrimination of the appropriate nucleotide label precisely to the correct respective sample molecule. Alternatively, said diverse affinity labels may be added to sample molecules so as to be independently recognizable by appropriate receptor molecules or other affinity means, each complementarity type of which is respectively labeled with some distinct independently perceptible label.

Such labeling methods permit the processing of samples in fluid flow based apparatus without the loss of single molecule identifiability or assignability of results. Also note that manipulation with a laser trap, as for instance described by T.T. Perkins et al., may be employed with such uniquely labeled molecules.

Note that a case of encoding of particular interest is that of a functional molecule coupled to an informational molecule which is sufficient to direct the synthesis of said functional molecule in an appropriate, (e.g. biological or biological derived) system. Libraries of polypeptides expressed on the surface of, for example, bacteriophages carrying genetic material specifying said polypeptides, have found great use in the *in vitro* selection of binding specificities. Encoding which may additionally direct synthesis may be availed in the affinity characterization and molecular evolution applications of the present invention. The communication of a synthesis directing informational molecule (favorably DNA or RNA) with the correspondingly synthesized one or more functional molecules (generally a polypeptide) may be effected by the *in vivo* coupling or otherwise compartmentally enforced unique one-to-one corresponding coupling of said informational and said functional molecule. A particularly convenient instance of such a molecules comprises the fused expression of said functional molecule or molecules as segments of the terminal proteins of the informational molecules (i.e. DNA) of various virus (e.g. adenovirus) or bacteriophage (e.g. PRD1 or phi29) genomes. Alternatively, said functional molecules may be fused with some molecule which associates in a specific manner with said terminal proteins, and which has sufficient opportunity during its *in vivo* synthesis, without or preferably with concurrent viral or bacteriophage replication, to associate with the terminal protein of the genomic material which determines the composition of said functional molecules, such that upon purification

or lysis functional molecules remain in communication with the genetic material that determines their composition. Because biosynthesis of functional and informational moieties may favorably occur within the confines of a single cell, cross-coupling of inappropriate molecules may be readily avoided. Alternatively, the communication between polypeptide and polynucleotide moieties may be effected with some intermediate snRNP or snRNP-like moiety, where such an intermediate moiety may be targeted on the one hand by an appropriate affinity characteristic of one or more polypeptides to which said functional molecules are fused, and on the other hand by a polynucleotide sequence complementary (according to appropriate rules for double-, triple- or quadruple- helix formation) with the polynucleotide moiety of said intermediate snRNP or snRNP-like moiety.

Such complexes comprising an intermediate snRNP or snRNP-like moiety may also favorably be formed within the confines of a single cell.

1.4.9.5 CYBERNETIC MOLECULAR EVOLUTION AND ALGORITHM MEDIATED CYBERNETIC MOLECULAR EVOLUTION OF PHENOGENOCOUPLES

Such polynucleotide-polypeptide chimera, or other molecule types comprising thus communicating and informationally corresponding chimera (e.g. where the polypeptide moiety has further been subjected to post-translational modification such as specific glycosylation and has been associated by some method to the respective genetic material determining its composition, for example by the sorting of individual cells carrying said genetic material in the form of a DNA vector with terminal proteins and expressing and processing said polypeptide, into distinct wells or vessels followed by disruption of membranes such that terminal proteins fused with peptides having affinity for the particular polypeptide of interest may come into contact with the processed polypeptide of interest, comprising a method for the molecular evolution of multiple-biopolymer containing macromolecules), which may be termed phenogenocouples, may be used as sample molecules with the broad methods of the present invention to effect the affinity characterization (including either or both equilibrium and kinetic characterization of molecular recognition including catalytic recognition and catalysis) of functional moieties and then the characterization and

transcription of informational moieties thus determined to be of interest. Where algorithms control such a process, cybernetic molecular evolution is embodied.

Selected informational molecules may be selectively replicated or transcribed by activatable (e.g. photodeprotectable and especially 3' hydroxyl photodeprotectable) primers with appropriate complementarity to some region which bounds the informational content specifying said functional molecule or molecules. Alternatively, immobilization of a sample to be subjected to such manipulations may be effected so as to comprise some photolabile linkage, which may then be subjected to selective photodegradation to effect specific release. For immobilized samples, informational molecules which carry the relevant genetic component of a phenogenocouple may thus be released by either of these methods either singly, or as the population of multiple such molecules simultaneously, copied or otherwise released according to the pattern of deprotection.

Alternatively, successive generations of molecules need only be related informationally, by analysis of composition of one generation, by, for example, the massively parallel characterization methods of the present invention, followed by de novo synthesis of molecules carrying the desired complexity and diversity of the succeeding generation. This is a particular distinguishing feature of cybernetic molecular evolution; selection, amplification and mutation may be directed strictly by algorithms which manipulate data gathered about one generation to determine the composition of a succeeding generation.

Released molecules may then be recovered for subsequent amplification, mutation and subsequent rounds of selection by similar or other methods, as will be obvious to those skilled in the art of in vitro molecular evolution.

Note that post transcriptionally modified polypeptide moieties or other phenogenocouples may also be selected and otherwise subjected to in vitro evolution by conventional means as well as by the massively parallel examination and modification methods of the present invention.

Because of the correspondence between the diversity generation and selection aspects of molecular evolution, and immunological recognition and memory, all of these methods may be directly applied to cybernetic immune system applications of the present invention.

Labeled reagents and signal amplification and elimination techniques:

The categories enumerated below are included for description and not limitation; other appropriate labeling methods will be obvious to those skilled in the arts of biotechnology, cell biology and cytology, microscopy, organic chemistry, biochemistry or recombinant DNA techniques.

Each category will comprehend a variety of specific variations, as will be obvious to those skilled in the relevant arts. Various labeling methods will generally correspond best to various detection methods.

1.4.10 DETECTION METHODS FOR THE PRESENT INVENTION

Non-radioactive labeling techniques have been explored and, in recent years, integrated into partly automated DNA sequencing procedures. These improvements utilize the Sanger sequencing strategy. The label (e.g. fluorescent dye) can be tagged to the primer (Smith et al., *Nature* M, 674-679 (1986) and EPO Patent No. 87300998.9; Du Pont De Nemours EPO Application No. 0359225; Ansorge et al., *J. Biochem. Biophys. Methods* 13, 325-32 (1986)) or to the chain-terminating dideoxynucleoside triphosphates (Prober et al. *Science* 218, 336-41 (1987); Applied Biosystems, PCT Application WO 91/05060). Based on either labeling the primer or the ddNTP, systems have been developed by Applied Biosystems (Smith et al., *Science* 235, G89 (1987); U. S. Patent Nos. 5 70973 and 689013), Du Pont De Nemours (Prober et al., *Science* 238, 336-341 (1987); U.S. Patents Nos. 881372 and 57566), Pharmacia-LKB (Ansorge et al., *Nucleic Acids Res.* 11, 4593-4602 (1987) and EMBL Patent Application DE P3 724442 and P3 805 808. 1) and Hitachi (JP I - 90844 and DE 4011991 AI). A somewhat similar approach was developed by Brumbaugh et al., (*Proc. Nat. Sci. US A* 85 5610-14 (1988) and U.S. Patent No. 4,729,947). An improved method for the Du Pont system using two electrophoretic lanes with two different specific labels per lane is described (PCT Application W092/02635). A different approach uses fluorescently labeled avidin and biotin labeled primers. Here, the sequencing ladders ending with biotin are reacted during electrophoresis with the labeled avidin which results in the detection of the individual sequencing bands (Brumbaugh et al., U.S. Patent No. 594676).

More recently even more sensitive non-radioactive labeling techniques for DNA using chemiluminescence triggerable and amplifiable by enzymes have been developed (Beck, OKeefe, Coull and Köster, *Nucleic Acids Res.* 12, 5115- 5123 (1989) and Beck and Köster, *Anal. Chem.* Q 2258-2270 (1990)). These labeling methods were combined with multiplex DNA sequencing (Church et al., *Science* 240, 185-188 (1988) and direct blotting electrophoresis (DBE) (Beck and Pohl, *EMBO J* Vol. 3: p 2905-2909 (1984)) to provide for a strategy aimed at high throughput DNA sequencing (Köster et al., *Nucleic Acids Res. Symposium Ser. No. 2,4*, 318- 321 (1991), University of Utah, PCT Application No. WO 90/15883). However, this strategy still suffers from the disadvantage of being very laborious and difficult to automate.

Multiple distinctly labeled primers can be used to discriminate sequencing patterns. For example, four differently labeled sequencing primers specific for the single termination reactions, e.g. with fluorescent dyes and online detection using laser excitation in an automated sequencing device. The use of eight differently labeled primers allow the discrimination of the sequencing pattern from both strands. Instead of labeled primers, labeled ddNTP may be used for detection, if separation of the sequencing fragments derived from both strand is provided, With one biotin labeled primer, sequencing fragments from one strand can be isolated for example via biotin-streptavidin coated magnetic beads. Possible is also the isolation via immunoaffinity chromatography in the case of a digoxigenin labeled primer or with affinity chromatography in case of complementary oligonucleotides bound to a solid support.

1.4.10.1 Fluorescent labels

In automated sequencing, fluorescence labeled DNA fragments are detected during migration through the sequencing gel by laser excitation. Fluorescence label is incorporated during the sequencing reaction via labeled primers or chain extending nucleotides (Smith, L. et. al., Fluorescence detection in automated DNA sequence analysis, Nature 321.674-89 1986), (Knight, P., Automated DNA sequencers, Biotechnology 6:1095-96 1988).

Detection methods for the present invention may favorably exploit fluorescent labeling techniques.

Genome sequencing applications of the present invention may thus avail of established fluorescent modification and detection methods. Other applications of the present invention may also benefit from the application of fluorescence modification and detection methods.

Much effort has already been invested in the development of fluorescently labeled nucleotide triphosphate compounds and analogs thereof. Many such compounds are acceptable substrates for polynucleotide polymerase molecules. These compounds have therefore proven suitable for use in various electrophoresis based DNA sequencing methodologies utilizing fluorescence detection, as well as in other applications such as chromatin mapping. There are therefore various compounds comprising a fluorescent dye moiety and a nucleotide triphosphate moiety commercially available.

Fluorescent labels find use in variety of different biological, chemical, medical and biotechnological applications. One example of where such labels find use is in polynucleotide sequencing, particularly in automated DNA sequencing, which is becoming of critical importance to large scale DNA sequencing projects, such as the Human Genome Project.

In methods of automated DNA sequencing, differently sized fluorescently labeled DNA fragments which terminate at each base in the sequence are enzymatically produced using the DNA to be sequenced as a template. Each group of fragments corresponding to termination at one of the four labeled bases are labeled with the same label. Thus, those fragments terminating in A are labeled with a first label, while those terminating in G, C and T are labeled with second, third and fourth labels respectively. The labeled fragments are then separated by size in an electrophoretic medium and an electropherogram is generated, from which the DNA sequence is determined.

As methods of automated DNA sequencing have become more advanced, of increasing interest is the use of sets of fluorescent labels in which all of the labels are excited at a common wavelength and yet emit one of four different detectable signals, one for each of the four different bases. Such labels provide for a number of advantages, including high fluorescence signals and the ability to electrophoretically separate all of the labeled fragments in a single lane of an electrophoretic medium which avoids problems associated with lane to lane mobility variation.

Although such sets of labels have been developed for use in automated DNA sequencing applications, heretofore the differently labeled members of such sets have each emitted at a different wavelength. Thus, conventional automated detection devices currently employed in methods in which all of the enzymatically produced fragments or primer extension products are separated in the same lane must be able to detect emitted fluorescent light at four different wavelengths. This requirement can prove to be an undesirable limitation. More specifically, carrying out sequencing on vast numbers of different DNA templates simultaneously increases the number of different fragments and corresponding labels required. At the same time, there is a need for a reduction in the complexity of the detection device, e.g. a device which can operate with light detection at only two wavelengths is preferable.

Sets of fluorescent labels, particularly sets of fluorescently labeled primers, and methods for their use in multi component analysis applications, particularly nucleic acid enzymatic sequencing applications, are provided. At least two of the label members of the set are energy transfer labels having a common donor and acceptor fluorophore separated by sufficiently different distances so that the two labels provide distinguishable fluorescent signals upon excitation at a common wavelength. In further describing the subject invention, the subject sets will first be described in greater detail followed by a discussion of methods for their use in multi component analysis applications.

Before the subject invention is further described, it is to be understood that the invention is not limited to the particular embodiments of the invention described below, as variations of the particular embodiments may be made and still fall within the scope of the appended claims. It is also to be understood that the terminology employed is for the purpose of describing particular embodiments, and is not intended to be limiting. Instead, the scope of the present invention will be established by the appended claims.

It must be noted that as used in this specification and the appended claims, the singular forms "a," "an" and "the" include plural reference unless the context clearly dictates otherwise. Unless defined otherwise all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs.

The subject sets of fluorescent labels comprise a plurality of different types of labels, wherein each type of label in a given set is capable of producing a distinguishable fluorescent signal from that of the other types of labels in different sets. Labels in the different sets generate different signals, preferably, though not necessarily upon excitation at a common excitation wavelength. For DNA sequencing applications, the subject sets will comprise at least 2 different types of labels, and may comprise 8 or more different types of labels, where for many applications the number of different types of labels in the set will not exceed 6, and will usually not exceed four, where at least two of the different types of labels are energy transfer labels sharing a common donor and acceptor fluorophore, as described in greater detail below. For other applications, such as fluorescence in situ hybridization (FISH), substantially more than 8 labels are ideal so that multiple targets can be analyzed.

The distinguishable signals generated by the "at least two energy transfer labels" will at least comprise the intensity of emitted light at one to two wavelengths. Preferably, the distinguishable signals produced by the "at least two energy transfer labels" will comprise distinguishable fluorescence emission patterns, which patterns are generated by plotting the intensity of emitted light from differently sized fragments at two wavelengths with respect to time as differently labeled fragments move relative to a detector, which patterns are known in the art as electropherograms. For analyses not based on electrophoresis, such as micro- array chip based assays, different targets tagged with a specific label can be differentiated from each other by the unique fluorescence patterns. For example, in one type of label of a set the intensity of emitted light at a first wavelength may be twice that of the intensity of emitted light at a second wavelength and in the second label the magnitude of the intensities of light emitted at the two wavelengths may be reversed, or light may be emitted at only one intensity. The different patterns are generated by varying the distance between the donor and acceptor. These patterns emitted from each of these labels are thus distinguishable.

The subject sets will comprise a plurality of different types of fluorescent labels, where at least two of the labels and usually all of the labels are energy transfer labels which comprise at least one acceptor fluorophore and at least one donor fluorophore in energy transfer relationship, where such labels may have more complex configurations, such as multiple donors and/or multiple acceptors, e.g. donor 1, acceptor 1 and acceptor 2. Critical to the subject sets is that at least two of the labels of the sets have common donor and acceptor fluorophores, where the only difference between the labels is the distance between these common acceptor and donor fluorophores. Thus, for sets of labels in which each label comprises a single donor and a single acceptor, at least one of the energy transfer labels will have a donor fluorophore and acceptor fluorophore in energy transfer relationship separated by a distance x and at least one of the energy transfer labels will comprise the same donor and acceptor fluorophores in energy transfer relationship separated by a different distance y , where the distances x and y are sufficiently different to provide for distinguishable fluorescence emission patterns upon excitation at a common wavelength, as described above.

In those sets comprising a third label having the same donor and acceptor fluorophores as the first and second label, the distance z between the donor and acceptor fluorophore will be sufficiently different from x and y to ensure that the third label is capable of providing a distinguishable fluorescence emission pattern from the first and second labels. Thus, in a particular set of labels, one may have a plurality of labels having the same donor and acceptor fluorophores, where the only difference among the labels is the distance between the donor and acceptor fluorophores. To ensure that different types of labels of a set having common donor and acceptor fluorophores yield distinguishable fluorescence emission patterns, the distances between the donor and acceptor fluorophores will differ by at least about 5 %, usually by at least about 10 % and more usually by at least about 20 % and will generally range from about 4 to 200 Å, usually from about 12 to 100 Å and more usually from about 15 to 80 Å, where the minimums in such distances are determined based on currently available detection devices and may be reduced as detection technology becomes more sensitive, therefore more distinct labels can be generated.

In one preferred embodiment, at least a portion of, up to and including all of, the labels of the subject sets will comprise a donor and acceptor fluorescer component in energy transfer relationship and covalently bonded to a spacer component, i.e. energy transfer labels. Thus, one could have a set of a plurality of labels in which only two of the labels comprise the above mentioned donor and acceptor fluorescer components and the remainder of the labels comprise a single fluorescer component. Preferably, however, all of the labels will comprise a donor and acceptor fluorescer component. Generally, for one donor and one acceptor ET systems, if a set comprises n types of energy transfer labels, the number of different types of acceptor fluorophores present in the energy transfer labels of the set will not exceed $n-1$. Thus, if the number of different types of energy transfer labels in the set is four, the number of different acceptor fluorophores in the set will not exceed 3, and will usually not exceed 2.

In other preferred embodiments, additional combinations of labels are possible. Thus, in a set of labels, two of the labels could be energy transfer labels sharing common donor and acceptor fluorophores separated by different distances and the remaining labels could be additional energy transfer labels with different donor and/or acceptor fluorophores, non-energy transfer fluorescent labels, and the like.

In the energy transfer labels of the subject sets, the spacer component to which the fluorescer components are covalently bound will typically be a polymeric chain or other chemical moiety capable of acting as a spacer for the donor and acceptor fluorophore components, such as a rigid chemical moiety, such as chemicals with cyclic ring or chain structures which can separate the donor and acceptor and which also can be incorporated with an active group for attaching to the targets to be analyzed, where the spacer component will generally be a polymeric chain, where the fluorescer components are covalently bonded through linking groups to monomeric units of the chain, where these monomeric units of the chain are separated by a plurality of monomeric units sufficient so that energy transfer can occur from the donor to acceptor fluorescer components. The polymeric chains will generally be either polynucleotides, analogues or mimetics thereof, or peptides, peptide analogues or mimetics thereof, e.g. peptoids. For polynucleotides, polynucleotide analogues or mimetics thereof, the polymeric chain will generally comprise sugar moieties which may or may not be covalently bonded to a heterocyclic nitrogenous base, e.g. adenine, guanine, cytosine, thymine, uracil etc., and are linked by a linking group. The sugar moieties will generally be five membered rings, e.g. ribose, or six membered rings, e.g. hexose, with five membered rings such as ribose being preferred. A number of different sugar linking groups may be employed, where illustrative linking groups include phosphodiester, phosphorothioate, methylene(methyl imino)(MMI), methophosphonate, phosphoramidate, guanidine, and the like. See Matteucci & Wagner, *Nature* (1996) Supp 84: 20-22. Peptide, peptide analogues and mimetics thereof suitable for use as the polymeric spacer include peptoids as described in WO 91/19735, the disclosure of which is herein incorporated by reference, where the individual monomeric units which are joined through amide bonds may or may not be bonded to a heterocyclic nitrogenous base, e.g. peptide nucleic acids. See Matteucci & Wagner *supra*. Generally, the polymeric spacer components of the subject labels will be peptide nucleic acid, polysugarphosphate as found in energy transfer cassettes as described in PCT/US96/13134, the disclosure of which is herein incorporated by reference, and polynucleotides as described in PCT/US95/01205, the disclosure of which is herein incorporated by reference.

Both the donor and acceptor fluorescer components of the subject labels will be covalently bonded to the spacer component, e.g. the polymeric spacer chain,

through a linking group. The linking group can be varied widely and is not critical to this invention. The linking groups may be aliphatic, alicyclic, aromatic or heterocyclic, or combinations thereof. Functionalities or heteroatoms which may be present in the linking group include oxygen, nitrogen, sulfur, or the like, where the heteroatom functionality which may be present is oxy, oxo, thio, thiono, amino, amido and the like. Any of a variety of the linking groups may be employed which do not interfere with the energy transfer and gel electrophoresis, which may include purines or pyrimidines, particularly uridine, thymidine, cytosine, where substitution will be at an annular member, particularly carbon, or a side chain, e.g. methyl in thymidine. The donor and/or fluorescer component may be bonded directly to a base or through a linking group of from 1 to 6, more usually from 1 to 3 atoms, particularly carbon atoms. The linking group may be saturated or unsaturated, usually having not more than about one site of aliphatic unsaturation.

Though not absolutely necessarily, generally for DNA sequencing applications at least one of the donor and acceptor fluorescer components will be linked to a terminus of the polymeric spacer chain, where usually the donor fluorescer component will be bonded to the terminus of the chain, and the acceptor fluorescer component bonded to a monomeric unit internal to the chain. For labels comprising polynucleotides, analogues or mimetics thereof as the polymeric chain, the donor fluorescer component will generally be at the 5' terminus of the polymeric chain and the acceptor fluorescer component will be bonded to the polymeric chain at a position 3' position to the 5' terminus of the chain. For other applications, such as FISH, a variety of labeling approaches are possible.

The donor fluorescer components will generally be compounds which absorb in the range of about 300 to 900 nm, usually in the range of about 350 to 800 nm, and are capable of transferring energy to the acceptor fluorescer component. The donor component will have a strong molar absorbance co-efficient at the desired excitation wavelength, desirably greater than about 10^4 preferably greater than about $10^5 \text{ cm}^{-1}\text{M}^{-1}$. The molecular weight of the donor component will usually be less than about 2.0 kD, more usually less than about 1.5 kD. A variety of compounds may be employed as donor fluorescer components, including fluorescein, phycoerythrin, BODIPY, DAPI, Indo-1, coumarin, dansyl, cyanine dyes, and the like. Specific donor compounds of interest include fluorescein, rhodamine, cyanine dyes and the like.

Although the donor and acceptor fluorescer component may be the same, e. g both may be FAM, where they are different the acceptor fluorescer moiety will generally absorb light at a wavelength which is usually at least 10 nm higher, more usually at least 20 nm or higher, than the maximum absorbance wavelength of the donor, and will have a fluorescence emission maximum at a wavelength ranging from about 400 to 900 nm. As with the donor component, the acceptor fluorescer component will have a molecular weight of less than about 2.0 kD, usually less than about 1.5 kD. Acceptor fluorescer moieties may be rhodamines, fluorescein derivatives, BODIPY and cyanine dyes and the like. Specific acceptor fluorescer moieties include FAM, JOE, TAM, ROX, BODIPY and cyanine dyes.

The distance between the donor and acceptor fluorescer components will be chosen to provide for energy transfer from the donor to acceptor fluorescer, where the efficiency of energy transfer will be from 20 to 100 %. Depending on the donor and acceptor fluorescer components, the distance between the two will generally range from 4 to 200 Å, usually from 12 to 100 Å and more usually from 15 to 80 Å, as described above.

For the most part the labels of the subject sets will be described by the following formula:



wherein: D is the donor fluorescer component, which may consist of more than two different donors separated by a spacer;

N is the spacer component, which may be a polymeric chain or rigid chemical moiety, where when N is a polymeric spacer that comprises nucleotides, analogues or mimetics thereof, the number of monomeric units in N will generally range from about 1 to 50, usually from about 4 to 20 and more usually from about 4 to 16;

A is the acceptor fluorescer component, which may consist of more than two different acceptors separated by a spacer; and X is optional and is generally present when the labels are incorporated into oligonucleotide primers, where X is a functionality, e.g an activated phosphate group, for linking to a mono- or

polynucleotide, analogue or mimetic thereof, particularly a deoxyribonucleotide, generally of from 1 to 50, more usually from 1 to 25 nucleotides.

For sets to be employed in nucleic acid enzymatic sequencing in which the labels are to be employed as primers, the labels of the subject sets will comprise either the donor and acceptor fluorescer components attached directly to a hybridizing polymeric backbone, e.g. a polynucleotide, peptide nucleic acid and the like, or the donor and acceptor fluorescer components will be present in an energy transfer cassette attached to a hybridizable component, where the energy transfer cassette comprises the fluorescer components attached to a non-hybridizing polymeric backbone, e.g. a universal spacer. See PCT/US96/13134 and Ju et al., Nat. Med. (1996) *supra*, the disclosures of which are herein incorporated by reference. The hybridizable component will typically comprise from about 8 to 40, more usually from about 8 to 25 nucleotides, where the hybridizable component will generally be complementary to various commercially available vector sequences such that during use, synthesis proceeds from the vector into the cloned sequence. The vectors may include single-stranded filamentous bacteriophage vectors, the bacteriophage lambda vector, pUC vectors, pGEM vectors, or the like. Conveniently, the primer may be derived from a universal primer, such as pUC/M13, g t10, g t11, and the like, (See Sambrook et al., Molecular Cloning: A Laboratory Manual., 2nd ed., CSHL, 1989, Section 13), where the universal primer will have been modified as described above, e.g. by either directly attaching the donor and acceptor fluorescer components to bases of the primer or by attaching an energy transfer cassette comprising the fluorescer components to the primer.

Sets of preferred energy transfer labels comprising donor and acceptor fluorescers covalently attached to a polynucleotide backbone in the above D-N-A format include: (1) F6R, F13R, F16R and F16F; where different formats can be employed as long as the four primers display distinct fluorescence emission patterns.

The fluorescent labels of the subject sets can be readily synthesized according to known methods, where the subject labels will generally be synthesized by oligomerizing monomeric units of the polymeric chain of the label, where certain of the monomeric units will be covalently attached to a fluorescer component.

The subject sets of fluorescent labels find use in applications where at least two components of a sample or mixture of components are to be distinguishably

detected. In such applications, the set will be combined with the sample comprising the to be detected components under conditions in which at least two of the components of the sample if present at all will be labeled with first and second labels of the set, where the first and second labels of the set comprise the same donor and acceptor fluorescer components which are separated by different distances. Thus, a first component of the sample is labeled with a first label of the set comprising donor and acceptor fluorescer components separated by a first distance X. A second component of the sample is labeled with a second label comprising the same donor and fluorescer components separated by a second distance Y, where X and Y are as described above. The labeled first and second components, which may or may not have been separated from the remaining components of the sample, are then irradiated by light at a wavelength capable of being absorbed by the donor fluorescer components, generally at a wavelength which is maximally absorbed by the donor fluorescer components. Irradiation of the labeled components results in the generation of distinguishable fluorescence emission patterns from the labeled components, a first fluorescence emission pattern generated by the first label and second pattern being attributable to the second label. The distinguishable fluorescence emission patterns are then detected. Applications in which the subject labels find use include a variety of multicomponent analysis applications in which fluorescent labels are employed, including FISH, micro-array chip based assays where the labels may be used as probes which specifically bind to target components, DNA sequencing where the labels may be present as primers, and the like.

The subject sets of labels find particular use in polynucleotide enzymatic sequencing applications, where four different sets of differently sized polynucleotide fragments terminating at a different base are generated (with the members of each set terminating at the same base) and one wishes to distinguish the sets of fragments from each other. In such applications, the sets will generally comprise four different labels which are capable of acting as primers for enzymatic extension, where at least two of the labels will be energy transfer labels comprising differently spaced common donor and acceptor fluorescer components that are capable of generating distinguishable fluorescence emission patterns upon excitation at a common wavelength of light. Using methods known in the art, a first set of primer extension products all ending in A will be generated by using a first of the labels of the set as a primer. Second, third

and fourth sets of primer extension products terminating in G, C and T will be also be enzymatically produced. The four different sets of primer extension products will then be combined and size separated, usually in an electrophoretic medium. The separated fragments will then be moved relative to a detector (where usually either the fragments or the detector will be stationary). The intensity of emitted light from each labeled fragment as it passes relative to the detector will be plotted as a function of time, i.e. an electropherogram will be produced. Since, the labels of the subject sets will generally emit light in only two wavelengths, the plotted electropherogram will comprise light emitted in two wavelengths. Each peak in the electropherogram will correspond to a particular type of primer extension product (i.e. A, G, C or T), where each peak will comprise one of four different fluorescence emission patterns. To determine the DNA sequence, the electropherogram will be read, with each different fluorescence emission pattern related to one of the four different bases in the DNA chain.

Where desired, two sets of labels according to the subject invention may be employed, where the distinguishable fluorescence emission patterns produced by the labels in the first set will comprise emissions at a first and second wavelength and the patterns produced by the second set of labels will comprise emissions at a third and fourth wavelength. By using two such sets in conjunction with one another, one could detect primer extension products produced from two different template DNA strands at essentially the same time in a conventional four color detector, thereby doubling the throughput of the detector.

The subject sets of labels may be sold in kits, where the kits may or may not comprise additional reagents or components necessary for the particular application in which the label set is to be employed. Thus, for sequencing applications, the subject sets may be sold in a kit which further comprises one or more of the additional requisite sequencing reagents, such as polymerase, nucleotides, dideoxynucleotides and the like.

The following examples are offered by way of illustration and not by way of limitation. The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the subject sets of fluorescent labels.

1.4.10.2 Affinity labels

Other single molecule detection methods have availed of compounds having well studied affinity interactions with other molecules, such as receptor- ligand interactions.

Genome sequencing applications of the present invention may thus avail of established affinity labeling and detection methods. Other applications of the present invention may also benefit from the application of affinity labeling and detection methods.

Various compounds comprising a nucleotide triphosphate moiety and a small molecule affinity moiety are commercially available and suitable as substrates for DNA polymerases. Said compounds have been used, in conjunction with DNA polymerases, to effect the affinity labeling of various polynucleotide molecules, and thus labeled polynucleotides are routinely subjected to manipulations comprising the formation of an affinity association with an appropriate receptor molecule. Two common examples are the use of biotin as said affinity moiety and streptavidin as said receptor molecule, and digoxigenin as said affinity moiety and anti- digoxigenin antibodies or fragments thereof as the respective said receptor molecule. It will be obvious to those skilled in the relevant arts that there are numerous other possible ligand-receptor interactions which may be exploited for affinity labeling purposes as well as immobilization purposes of the present invention, and that multiple distinct affinity interactions may be employed simultaneously.

For detection purposes, said affinity labels may be used to bind a microscopic colloid or bead which has been modified with an appropriate complementary affinity group such as a receptor.

1.4.10.1 Affinity Label Detection With Microscopic Beads

In recent years a number of different methods and materials have been developed to permit the affinity binding of beads to molecules. Such binding is commonly accomplished by coating said beads with receptor molecules, such as streptavidin or Protein A (also known as Staph A, to which immunoglobulin G antibodies may subsequently be bound). Bead types include polymeric spheres of micron or submicron dimensions, metallic colloids such as colloidal gold, silica beads

and magnetic beads. As will be obvious to those skilled in the art of polymer chemistry, polymer beads including dendrimers may incorporate dyes or liquid crystal molecules as side chains or within polymeric backbones, and these may facilitate optical detection methods. Attachment of appropriate receptor or affinity molecules to the surfaces of such beads yields a reagent suitable for the detection of an affinity labeled molecule. One such detection scheme was utilized by Finzi and Gelles, albeit for different purposes.

1.4.10.3 Multimeric labels:

Where sensitivity to a single labeling moiety is insufficient, labeled reagents may comprise multiple occurrences of said labeling moiety in a manner that does not interfere with the corresponding molecular recognition and monomer addition processes, to increase the likelihood of correct signal amplification of any labeled molecule. For example, the ordinary single biotin moiety attached to a nucleotide by a linker may be replaced with a polymer having multiple biotin moieties as side chains, such that the likelihood of a streptavidin molecule interacting with this multimeric affinity label is increased. Fluorescent labels may similarly be multiplied, as may any other labeling moieties. Measures must be taken in the design and synthesis of such multimerically labeled reagents to ensure that solubility is retained. This may be accomplished by choosing a highly soluble polymer as the backbone carrying said labeling moieties comprising the multimeric label.

1.4.10.4 Polymerization nucleating labels:

Any compound capable of serving as an initiator for some aqueous polymerization may also serve as a labeling moiety. This initiator nucleates the formation of a perceptible polymer attached to the sample molecule. Such a polymer, may, for example, comprise multiple fluorescent moieties, or simply effect a local change in transmittance of light or a local change of refractive index. After detection has been accomplished, said perceptible polymer is degraded or otherwise removed from the sample molecule. Such polymerizations may be self-limiting, as is the case for some dendrimeric polymers.

For this label detection methodology, polymerization is caused to occur in a step after the labeled nucleotide is added to the sample molecule, and must proceed via a chemistry that leaves the sample molecule intact. Degradation or removal of said perceptible polymer must also leave the sample molecule intact. Subject to the

above stated limitation, any polymer and respective detection method may be employed.

1.4.10.5 Enzymatic labels and conjugates thereof

1.4.10.5.1 Photochemical labeling

Various methods have been developed for the photochemical labeling of molecules and especially biological macromolecules. These include detection of affinity labels such as biotin with conjugates of streptavidin and an appropriate enzyme capable of catalysing the formation of a chromophore from a chromophorigenic substrate, or capable of catalysing a photon liberating chemical reaction, as with the enzyme luciferase. Such photochemical labeling methods will be readily applicable as detection methods for various embodiments of the present invention.

Note that multimeric affinity labels accessible for simultaneous association with multiple such enzymes will enable greater signal amplification, as will secondary enzyme amplification techniques and other techniques known within the molecular biological and microscopic arts.

1.4.10.5.2 Cleavable linkers

Labeling moieties are favorably in communication with or coupled to nucleotides via a linker of sufficient length to ensure that the presence of said labeling moieties on said nucleotides will not interfere with the action of a polymerase enzyme on said nucleotides. Linkers will also necessarily be of some minimal length when stepping control is effected through the use of various preformed enzyme-nucleotide complexes (as described below). Once a nucleotide has been added by polymerization to (the daughter strand of) a sample molecule, and the accompanying label has been detected, proper detection and discrimination of subsequent nucleotides requires the elimination of said accompanying label. This may favorably be accomplished through the cleavage of said linkers which have been designed and synthesized to admit of cleavage by treatments which will not degrade or otherwise modify the relevant state or information content of sample molecules.

Cleavability may be provided for in a number of ways which will be obvious to those skilled in the arts of organic and synthetic chemistry. For example, said linker may include along its length one or more ester linkages, which will be susceptible to hydrolysis, which may be sufficiently mild for various ester functional

groups. Amide linkages may similarly be employed. Linkages comprising disulfide bonds within their length have been developed to provide for cleavability; reagents comprising such linkages are commercially available and have been used to modify nucleotides in a manner which may be conveniently reversed by treatment with mild reducing agents such as dithiothreitol. Cleavable linkages may be provided so as to minimize the portion of the linker which remains on the sample molecule. Because polymerases are relatively tolerant of linkers which may extend from various atoms of nucleotide molecules, it is not, however, critical that all of said linkers be cleaved away from the nucleotides incorporated into said sample molecules in the process of label removal.

Note that commercially available biotin derived nucleotides frequently contain, along the linker joining said biotin moiety to said nucleotide moiety, one or more ester or amide bonds, which is susceptible to cleavage by various chemical treatments.

Note also that for linkers comprising appropriate bonds along their length, enzymatic cleavage may be performed.

1.4.10.5.3 Dissociative cleavage:

Note that cleavage of a labeling moiety may also be effected by the disruption of some affinity interaction which effects the communication between said labeling moiety and the nucleotide moiety. In such cases, moieties joined by non-covalent associations may, for example, be dissociated by physical or chemical changes which do not necessarily cleave covalent bonds.

Photocleavable moieties may also comprise an intermediate portion of linkers joining labeling moieties to nucleotide moieties, such that upon photocleavage of said photocleavable moieties, communication between the termini of said linker is disrupted and the label moiety is liberated from the nucleotide moiety. Because photodeprotection or photocleavage reactions generally proceed quite rapidly, with appropriate detection and photoexcitation means, detection, label removal and nucleotide incorporation rates per sample molecule may approach the limit imposed by any particular polymerase enzyme and the processivity of said enzyme. Long linkers with photocleavable termini have been synthesized.

Similarly, compounds which thermally degrade into two or more portions may comprise an intermediate portion of such linkers, such that thermal cycling may

be employed to effect linker cleavage. Thermostable polymerases may be conveniently employed in embodiments availing thermolabile linkers.

1.4.10.5.4 Photomodification

Single dye molecule photobleaching has been directly observed. Fluorescent labels of nucleotides, particularly when only one or a small number of such moieties are used for labeling, may be neutralized by photobleaching, such that while some product of said fluorescent label may remain in communication with the sample molecule (e.g. the daughter strand of a polynucleotide being sequenced) it will no longer provide a signal sufficiently strong to interfere with the detection and discrimination of subsequently added labels.

Beyond photobleaching of fluorescent labels, affinity labels with appropriate photochemical properties may be subjected to photochemical modification rendering them inert to binding, generally subsequent to dissociation of the corresponding receptor by appropriate means.

For affinity labels, fluorescent labels or any other labeling moieties, chemical modification appropriate to the chemistries of said labels which effects a change or reduction in the detectable signal provided by said label may be availed to prevent interference of said labels with similar or distinct labels subsequently added to sample molecules or complexes thereof.

1.4.10.5.5 Labeling with activation and thermodynamic decay:

Compounds such as spirobenzopyran, which have labile, structurally and photochemically distinct but interconvertible isomers, may be used as labeling moieties. Here, an excited state of such a moiety may be used as a means of detection. After said detection has been successfully effected, chemical modification of one or another state of such interconvertible molecules may then neutralize it. Alternatively, activation may cause such a label to convert to some unstable but discernible state, which then irreversibly degrades according to characterizable kinetics. Such molecules must be chosen so as to remain in said discernible state for a sufficient time period to permit detection, but reliably degrade (to completion for a population of such molecules) within a practical time period.

1.4.10.5.6 Binding reaction inhibition detection methods:

Agents which specifically inhibit binding reactions may be identified rapidly through the detection of molecules, of a diverse library each molecular species of

which is uniquely labeled, not bound by particles some sample which may comprise many different species, in the presence test reagent, which is labeled, and permitted to associate with said sample' (preferably during a preincubation step before the addition of said diverse library to said sample,) in analogy to blocking antibody assays. Results are compared to those obtained with an aliquot of said diverse library and another portion of the same said sample. Such an assay may be performed for increasing concentrations of said test reagent.

1.4.10.5.7 Enzymatically enforced associations at defined molecular sites:

Methods are provided to enforce highly specific associations and reactions, including molecular recognition processes, on individual sample molecules or on populations and subpopulations of sample molecules. These are described for genome sequencing applications, but the methods included thereunder have broad applicability, including to any molecular affinity interaction.

1.4.10.5.8 Enzymatically enforced template directed copolymer addition at defined site:

Controlled comonomer addition Various methods may be used to accomplish the controlled addition of monomers, including nucleotides and especially labeled or protected nucleotides, to the daughter strand of a sample template molecule.

1.4.10.6 Rate control or accommodation:

Means of slowing the time required for the addition of a single nucleotide to a sample molecule will circumvent the requirement of stepping control. This will be particularly applicable for detection mechanisms not requiring separate manipulation steps (such as the separate association of beads to affinity labeled sample molecules). For example, the four nucleotides, each respectively labeled with unique, removable or neutralizable fluorescent labels, may be added to appropriately primed sample template molecules in the presence of polymerases, at low concentrations. Said concentrations must be sufficiently low that two nucleotides are not added to the sample molecule in less than the time required to accomplish the detection of the first such addition. Because all labels are present in the observation field, detection is accomplished through the observation of the reduction of the Brownian motion of a fluorescent moiety due to its addition to the sample molecule, in close analogy to the experiments of Finzi and Gelles, but it will be noted that the change in mobility is much larger in the present case. Alternatively detection may be understood to depend

on an increase in the net residence of some fluorescent moiety within a defined region or the occupancy of said a region, above the occupancy arising from the background of unbound labeled nucleotides.

Such detection is preferably conducted with a scanning excitation beat fluorescence confocal microscopic method as described above, or with a scanning detection light path, as also described above. Conditions (particularly nucleotide concentration) are chosen such that on average less than one labeled nucleotide will be present within the area illuminated by such a beam or thus observed, so that a light pulse of appropriate frequency passing through, for example, the pinhole which effects the scanning of the excitation beam, may be used to photobleach or photocleave the fluorescent label from the sample molecule after it has been detected to have been added to the sample molecule, without the appreciable accumulation of incidentally unlabeled nucleotides. Alternatively, an SLM may be used to spatially control illumination of the sample by an appropriate frequency of light to effect photochemical unlabeled, and thus permit the simultaneous unlabeled of multiple sample molecules.

This method may be understood as concentration modulated control of the kinetics of polymerization processivity, which is used to facilitate direct observation of successive addition of individual (labeled) nucleotides, with controlled unlabeled. Scanning rate and other instrumentation dependent parameters will influence optimal conditions and concentrations. Thus, direct observation of the addition of comonomers is dynamically observed, and sequence information for the respective sample molecule may be reconstructed accordingly.

1.4.10.7 Stepping control by equilibrium means:

A simple method to effect adequate stepping control for sequencing applications of the present invention relies on equilibrium control. In this method, nucleotides (which are labeled) are limiting, and there is a relative excess of sample molecules. Exonuclease activity intrinsic to most polynucleotide polymerases is circumvented by the use of alpha- phosphorothioate nucleotides (which are appropriately labeled) which are resistant to such degradation, in this method. Other nucleotide derivatives or analogs suitable as substrates for polymerases and yielding exonuclease resistant polynucleotides may likewise be employed.

As an example of equilibrium controlled stepping, a thirty-three-fold excess of sample molecules relative to labeled complementary nucleotides per cycle may be chosen. Polymerase molecules are preferably provided in excess of sample template molecules. Each sample molecule has a three percent chance of undergoing a single nucleotide addition. Nucleotides are rapidly depleted. Any sample molecule which has undergone one nucleotide addition has a further three percent chance, or in total approximately a 0.1% chance of undergoing a second nucleotide addition. For a sequencing segment run of 20 bases per sample molecule, each segment will experience an error contribution of $(20)(0.1\%)$ or 2% from multiple additions within a cycle. Such erroneous segment data will be conspicuous when oversampling is performed due to the correspondingly low frequency with which it occurs.

Alternatively, for tenfold excess of sample molecules with respect to labeled complementary nucleotides, there is a 1% chance per base of multiple additions to the same molecule, or, again for sequencing runs of bases, a 20% chance that a segment experiences at least one duplicate addition event. For five-fold oversampling, the binomial distribution indicates that there is approximately a 94.2% chance that three or more segments including a particular base contain correct data regarding that base. Any specific individual data error is highly unlikely to occur more than once for fivefold oversampling. Note that in practice such calculations will also have to account for label amplification error and label detection error, but these error contributions should be susceptible to reduction to manageably low levels.

More generally, for a ratio x of nucleotide molecules to sample template molecules with a complementary base properly located relative to the primer, for $x < 1$ there is a probability p equal to x that a particular sample molecule will experience the addition of at least one nucleotide and a probability p^k that any sample molecule will experience at least k nucleotide additions within the same sequencing cycle. Multiple nucleotide additions to a sample molecule within the same sequencing cycle will result in erroneous sequence information being obtained from said sample molecule. The probability (d) of such a multiple incorporation error occurring within the sequence segment data obtained from a particular sample molecule in a sequencing run of n bases will be less than $2(n)(p^2)$. The net sequence information per sample molecule obtained per sequencing cycle will be x bases, and the net sequence information for a sample with N molecules will be $(x)(N)$ bases,

which will be large for large N. For example, with $x=.03$ and $N=3.3 \times 10^{10}$, there will be a net raw data accumulation of approximately 10^9 bases per cycle, which, with one-hundred-fold oversampling (i.e. due to each sequence being represented 100 times in the sample) will yield 10^7 bases of data per cycle; for a desired segment length of $n=15$ bases, $n/x=(15)/(.03)$ or approximately 500 sequencing cycles will be required per run, and the run will yield 1.5×10^8 bases of information. For polymerase fidelity of 95% (an extremely low value chosen for purposes of illustration) there will be a 5% error rate (e) per base or a segment error rate of $(n)(e)=75\%$ per molecule, but the probability of two erroneous sequence segments having identical sequences will be $e2(1-e)^{n-1}$ for segments with a single base error, which will be the most frequent error species. For this example, this yields a 0.12% frequency. Methods similar to those used to determine consensus sequences may thus be employed to obtain highly accurate data in spite of less than perfect polymerization fidelity. Thus, fidelity error components will be negligible compared to multiple base incorporation errors. For this example, multiple base incorporation error components will yield an error rate of less than $(2)(15)(.03)^2$ or about 3% per molecule. Again, oversampling will readily detect such errors, which will occur identically for two molecules with only $d2=(.03)^2$ or less than 0.1% probability, yielding a far lower error rate for over sampled data.

1.4.10.8 Stepping control by removable protecting groups:

Stepping control may favorably be applied to any polymerization process useful within the scope of the present invention, including both genome sequencing and affinity characterization applications.

Template directed polymerization depends on the processive addition of comonomers at the terminus of a growing daughter strand as specified by the respective complementary base of the parent template strand. Complementarity may be enforced through molecular recognition of said complementarity of protected analogs of said comonomers with the appropriate base of a template molecule, by the action upon such protected comonomers of appropriate polymerase enzymes.

Numerous monomers which may thus be added but do not provide an appropriate chemical functional group for subsequent elongation of the polynucleotide strand to which they have been enzymatically added are known within the relevant arts, and are generally referred to as chain terminators. Any such terminators which may be chemically or photochemically modified, particularly in a

manner not disrupting the sample molecule, to a form which may support subsequent addition of comonomers in the usual manner, may be employed to effect controlled stepping of polymerization addition.

Removable protecting groups are particularly advantageous for the genome sequencing applications of the present invention because they may be utilized to permit and ensure that exactly one nucleotide is added to a sample molecule per sequencing cycle. This will permit an even greater rate of data accumulation than may be achieved by equilibrium control methods, with which only a fraction of the sample molecule population per cycle yields data.

Photoremovable protecting groups may be used to gain similar advantage but further permit controlled spatial localization of deprotection. Examples of such nucleosides have been prepared.³¹ Because photodeprotection reactions generally proceed rapidly, with appropriate detection and photoexcitation means, processivity and nucleotide incorporation rates per sample molecule may approach the limit imposed by any particular polymerase enzyme.

Nucleotide analogs comprising such removable protecting groups preferably further comprise labeling moieties. A particularly convenient category of such compounds comprises a labeling moiety or multimer thereof in communication with the nucleotide moiety exclusively through said removable protecting group. For such compounds, removal of said removable protecting group will simultaneously effect removal of said labeling moiety. Simultaneous removal of both protecting moiety and labeling moiety will conveniently prepare a sample molecule for the next sequencing cycle in a single step.

Enzymological evidence concerning binding of 3' acetate esterified nucleotides and 5'-triphosphate-3'-(nucleoside-5'-monophosphate) to the triphosphate binding site of E. coli Polymerase I supports the acceptability of 3' modified nucleotides as substrates for this enzyme. Such protecting groups should therefore be compatible with either naturally occurring or genetically modified polymerases.

Note that in other applications of the present invention, primers comprising a photodeprotectable 3' hydroxyl terminus (which may be synthesized by the polymerization of an appropriate 3' protected nucleotide onto the unprotected 3' hydroxyl of an oligo- or poly-nucleotide, for instance, by the action of terminal deoxynucleotidyl transferase) may provide for the selective polymerization of a

polynucleotide moiety selectable by control over illumination of the appropriate region of the sample. A polynucleotide moiety to which such a primer is hybridized and then selectively deprotected may thus be subjected to amplification techniques such as PCR in a selectable manner. Such modified primers shall simply be referred to as photoactivatable primers.

The 3' deprotectable nucleotides employed in some variations on the present invention may also find other uses in molecular biology and biotechnology. They may be used as chain terminators in conventional enzymatic sequencing methods. If such manipulations are performed, any species terminating in a particular base may be extracted from the resolution medium (conventionally polyacrylamide gel), deprotected and then subjected to other manipulations requiring an active 3' hydroxyl group, such as ligation.

1.4.10.9 Enzyme adaptation to specific substrates:

The emergence of resistance to chain terminating nucleotide analogs by various viral polynucleotide polymerases suggests a convenient method for the in vitro evolution of polymerases capable of using reversibly 3' protected nucleotide analogs, or nucleotide analogs which otherwise serve as chain terminators which may be reactively modified to form an elongation competent molecule after incorporation into a polynucleotide. Further selection constraints may be concurrently or subsequently applied to fidelity, as the inclusion of non-sense codons in the coding region of a dominant lethal protein coding gene which is carried by the same genetic material carrying the polymerase gene under selection, such that misreading of the non-sense codon, by the polymerase under selection, will effect lethality to the host and thus select against low-fidelity polymerases.

As stated above, such deprotectable compounds may serve as a convenient stepping control means for polymerization. Included among such deprotectable nucleotides are nucleotides with photocleavable protecting groups, including those which reside on the 3' hydroxyl of a nucleotide.

1.4.10.10 Label encoding and labeling methods for data collection:

Various systems may be used to represent the data corresponding to the occurrence of an affinity interaction. The complexity required of such a representational system will be determined by the types of molecules and associations being examined and the extent to which manipulative steps are to be minimized.

The most rudimentary encoding system will be a one-bit binary labeling system, consisting of only one label moiety type, indicating whether or not an association of only one resolvable type occurred during the preceding association step.

For example, consider a sequencing application employing only a single nucleotide labeling moiety. Such a system may avail each of the four nucleotides modified with a biotin moiety attached by a sufficiently long, cleavable linker arm. In such a case, a polymerization sub-cycle comprises: the incubation of sample template molecules bearing appropriate primers with an appropriate polymerase and limiting quantities of only one labeled nucleotide (and no unlabeled nucleotides) such that this monomer will be added only if the template molecule has the complementary base in the template position immediately 5' to the base opposite the 3' terminal base of the

primer, and no monomers will be added otherwise; sample molecules are then washed to remove any remaining free nucleotides; the sample is then exposed to excess quantities of streptavidin modified fluorescently labeled beads for a sufficient length of time to ensure that all biotin moieties are bound by said labeled beads, and then all unbound beads are washed away; detection is then performed and data recorded; linkers are then cleaved. Said sub-cycle is repeated for the remaining three nucleotides, to constitute a cycle which successively tests for tile presence in the sample template molecule of each type of base immediately to the base opposite the 3' terminal base of the primer. If a sample molecule does not bind any label through such a cycle, then it was most likely "missed" due to the limiting concentration of nucleotides used to effect stepping of polymerization. If a sample molecule is labeled multiply during such a cycle, then the respective subsequent bases are detected as occurring in the template according to the pattern of labeling.

A somewhat more efficient encoding system is provided if two distinct labeling moieties may be availed. Each nucleotide will be indicated by the presence or absence of each of the two moiety types, as a binary code. The moieties may, for instance be biotin (B) and digoxigenin (D). For example, the representation may be: A=B+D; T=B; G=D. These three nucleotides are added for a first polymerization sub-cycle, and all unbound reagents then washed away. Either two perceptibly distinct bead types may be used for simultaneous detection, provided distinct affinity labels are sufficiently well separated by extended linkers for simultaneous binding, or a single bead type with two distinct receptor molecules may be used in two separate binding and release cycles, in which case the release of one bead type will have to leave the remaining affinity moiety bound to sample molecules.

After detection of bead labels, all remaining beads are removed and a second subcycle with C nucleotides affinity labeled with only one moiety are then polymerized onto sample molecules and appropriate detection is performed. Where protecting groups are used to effect stepping control, only one sub-cycle is needed and C may be unlabeled. In such cases unlabeled molecules will be detected as having added a cytidine.

More conveniently, nucleotides of each of the four types distinctly labeled with a fluorescent dye moiety may be used with fluorescence detection means, and a sequencing cycle consisting of only one sub-cycle. Alternatively, four antibodies (or

four other appropriate receptor molecules or affinity reagents) which each bind each of the four distinct dye moieties may be bound to each of four perceptibly distinct beads. In another arrangement, nucleotides may each be labeled with some distinct combination of multiple dye moieties, again encoding a unique binary label.

1.4.11 UTILITY OF THE SEQUENCE OF A GENOME

The present invention provides methods of detection and discrimination which address the complexity found in biological systems, though they may further be applied to non-natural systems including but not limited to mimetics. Much of this complexity derives from combinations or permutations of simple units such as the four nucleotide bases of polydeoxyribonucleic acids and polyribonucleic acids, or the twenty common amino acids found in polypeptides and proteins.

This complexity, which underlies the most diverse and nuanced of biological processes, has presented both the promise that ultimately much mechanistic knowledge of biological processes may be gained through the accumulation of greater information about underlying structures and biopolymer sequences, and the correspondingly motivated challenge of full enumeration and determination of these structures and sequences.

Because typical eukarvotic genomes contain between 10^7 and 10^{10} DNA base pairs, and because there are several well studied organisms of particular interest, economical and technically simple methods capable of determining the full genome sequence of an individual organism over a convenient~ short period of time would be particularly desirable.

The present invention can find applications in many fields, for instance, medical, diagnostic, forensic, genetics, biotechnology, and genome research. It should be noted that this technique would be applicable in many other fields and instances, and such applications would be discernible by people of ordinary skills in the respective fields.

The availability of such sequencing methods would enable greater clinical applications of molecular medicine, would facilitate greater and safer application of gene therapy, would permit timely completion of the several genome projects within fiscal constraints, and would enable facile gathering of genome information on populations of individuals, which would have applications in such areas as the study of polygenic diseases, epidemiology and field ecology. Such applications are presently limited by the cost and cumbersome nature of existing sequencing methodologies.

Combinatorial chemistry, affinity characterization, therapeutic synthetic immunochemistry, pharmacology and drug development, in vitro evolution and other

fields concerned with the elaboration of a diverse population of molecules, their characterization according to desired properties, and recovery or identification of molecules displaying suitable characteristics may be favorably improved by the availability of methods which permit the introduction of and both qualitative and quantitative characterization of kinetic and equilibrium properties of molecular recognition and binding phenomena, particularly where such parameters may be used as selective constraints.

There has further been some interest in rebuilding or supplementing the immune systems of immunocompromized individuals, and in the development of highly specific antibiotic agents targeted to antibiotic, antifungal or antiviral resistant or otherwise poorly treatable pathogens. Both of these goals may be furthered by the use of the methods of the present invention as they may readily be applied to the determination of pathogen specificity and antigenicity.

1.4.11.1. Application: Gene finding

An integrated clone map is constructed by the method described herein. When the bin probes include polymorphic genetic markers, and these markers are typed against the DNAs of member of families carrying a genetic trait, that trait can be genetically localized on the map relative to one or more bin probes. Depending on the study design, this genetic localization can be carried out using one of a variety of methods (G. M. Lathrop and J.-M. Lalouel, "Efficient computations in multilocus linkage analysis," *Amer. J. Hum. Genet.*, vol. 42, pp. 498-505, 1988; T. C. Matise, M. W. Perlin, and A. Chakravarti, "Automated construction of genetic linkage maps using an expert system (MultiMap): application to 1268 human microsatellite markers," *Nature Genetics*, vol. 6, no. 4, pp. 384-390, 1994; E. S. Lander and D. Botstein, "Mapping Complex Genetic Traits in Humans: New Methods Using a Complete RFLP Linkage Map," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. LI, Cold Spring Harbor, Cold Spring Harbor Laboratory, 1986, pp. 49-62; L. Penrose, *Ann. Eugenics*, vol. 18, pp. 120-124, 1953; N. E. Morton, *Am. J. Hum. Genet.*, vol. 35, pp. 201-213, 1983; N. Risch, *Am. J. Hum. Genet.*, vol. 40, pp. 1-14, 1987; E. Lander and D. Botstein, *Genetics*, vol. 121, pp. 185-199, 1989; N. Risch, "Linkage strategies for genetically complex traits," in three parts, *Am. J. Hum. Genet.*, vol. 46, pp. 222-253, 1990; N. Risch, *Genet. Epidemiol.*, vol. 7, pp. 3-16,

1990; N. Risch, *Am. J. Hum. Genet.*, vol. 48, pp. 1058-1064, 1991; P. Holmans, "Asymptotic Properties of Affected-Sib-Pair Linkage Analysis," *Am. J. Hum. Genet.*, vol. 52, pp. 362-374, 1993; N. Risch, S. Ghosh, and J. A. Todd, "Statistical Evaluation of Multiple-Locus Linkage Data in Experimental Species and Its Relevance to Human Studies: Application to Nonobese Diabetic (NOD) Mouse and Human Insulin-dependent Diabetes Mellitus (IDDM)," *Am. J. Hum. Genet.*, vol. 53, pp. 702-714, 1993; R. C. Elston, in *Genetic Approaches So Mental Disorders*, E. S. Gershon and C. R. Cloninger, ed. Washington DC: American Psychiatric Press, 1994, pp. 3-21), incorporated by reference.

Following genetic localization relative to the bin probes, the integrated contiged clone map provides an immediate means to proceed with positional cloning procedures. (D. Cohen, I. Chumakov, and J. Weissenbach, *Nature*, vol. 366, pp. 698-701, 1993; B.-S. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui, "Identification of the cystic fibrosis gene: genetic analysis," *Science*, vol. 245, pp. 1073-1080, 1989; J. R. Riordan, J. M. Rommens, B.-S. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Playisc, J.-L. Chou, M. L. Drumm, M. C. Iannuzzi, F. S. Collins, and L.-C. Tsui, "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA," *Science*, vol. 245, pp. 1066-1073, 1989), incorporated by reference. When an expression of candidate genes is included in the mapping resource (e.g., ESTs, cDNAs), the search may proceed more rapidly. When the genome sequences of the clones in the region have been determined, the gene search may be done in part using computer searches for candidate genes.

1.4.11.2 Application: Structure/function relation

The sequence of a genome is determined by the method described herein. From this genome sequence, the relation of a gene or its promoters to other known functions may be determined using similarity or homology searches. Protocols for these determinations are well described (N. J. Dracopoli, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, D. T. Moir, and D. Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995), incorporated by reference. The use of expressed sequence tag (EST) databases (Merck Gene Index, St. Louis, Mo.; Human Genome Sciences, Gathersburg, Md.) together with the

genome sequence provides a highly effective means for rapidly correlating a gene's sequence with the structure and function of its protein products.

1.4.11.3 Application: Metabolic network determination

The sequence of a genome is determined by the method described herein. Using the RT-PCR technique of differential display, perturbations on the cell state can be assayed in terms of DNA expression. Select perturbations can elucidate the metabolic networks of coupled enzyme systems in the cell. Reference back to the DNA sequence of the genome provides information about local control and gene/promoter interactions. This information can be used to understand disease mechanisms and to develop new pharmaceutical agent to alleviate said diseases.

1.4.11.4 Application: Growth and development

The sequence of a genome is determined by the method described herein. A method is described for constructing an integrated genetic-physical-expression map that includes the genome sequence and cDNAs. It is currently impractical to map very large numbers of cDNAs at high resolution, due to the currently used technology of sequencing each cDNA, constructing PCR primers for it, and then performing multiple PCR amplifications and detections relative to a panel of RHs to accurately map even a single cDNA. However unobvious it may currently seem to those skilled in the art, it would nonetheless be extremely desirable for elucidating the mechanism of cell growth and organism development to construct and map tissue-specific cDNA expression libraries at numerous points (e.g., at least every 24 hours) early in organism development. Further, the mapping of these expressed sequences back to their genomic locations would provide information on candidate genes, local gene expression, the coordination of normal and diseased cellular function under genetic control, and the time course of development in different tissues that would be highly useful in developing new diagnostic tests and therapeutic treatments for human disease. The method of the said examples provides such a novel means for practical rapid and high-resolution mapping of many expression libraries that would otherwise be neither constructed nor mapped.

1.4.11.5 Application: Drug development

A sequence and map of a genome is determined by the method described herein. The sequence of the human genome or integrated clone maps can be used to identify genes that are causative for human disease. From such genes, and their DNA promoters and protein products, mechanisms of diseases related to said genes can be determined. Pharmacological agents that intervene at key junctures in gene-related functions can then be devised to specifically circumvent and treat diseases related to these genes.

1.4.11.6 Application: Diagnostic testing

A sequence and map of a genome is determined by the method described herein. The sequence of the human genome or integrated clone maps can be used to identify genes that are causative for human disease. From such genes, and their DNA promoters and protein products, mechanisms of diseases related to said genes can be determined. Diagnostic tests that detect key junctures in gene-related structures and functions can then be devised to diagnose diseases related to these genes, and develop kits.

1.4.11.7 Application: Animal models

The sequence of a genome is determined by the method described herein. In the current art, sequencing even one complete mammalian genome is a highly debated and very expensive proposition (estimated to cost around one billion dollars) which is not likely to be performed more than once. However, the novel sequencing method described renders sequencing more practical, since it produces a high-resolution clone map which can be used to cost-effectively direct the sequencing effort and to practically assemble the resulting sequences. Given the pressing medical need for sequencing a mammalian genome, and the absence of any such useful coordinating map, clearly the described invention is highly nonobvious.

By constructing a map as described in the method described herein, the upfront burden of building maps for mammalian species other than humans is considerably reduced. Further, since the cost per base of sequencing is expected to diminish, particularly as newer sequencing technologies become established, the described method provides the first useful starting point for beginning (and eventually completing) the DNA sequence determination of model animal genomes. Comparison

of the DNA sequences and genes between human and model organisms is a well-established route for understanding and treating human disease.

1.4.11.8 Application: Somatic cell hybrids

The method described herein describes an inner product mapping analysis mechanism. Localization profiles are produced that can localize DNA sequences to high resolution. This inner product operation can be applied to somatic cell hybrid deletion panel data, thereby increasing the utility of such data by providing more confident and higher resolution localizations.

1.4.11.9 Application: Genome mismatch scanning

Genome mismatch scanning (GMS) (S. F. Nelson, J. H. McCusker, M. A. Sander, Y. Kee, P. Modrich, and P. O. Brown, "Genomic mismatch scanning: a new approach to genetic linkage mapping," *Nature Genetics*, vol. 4, no. May, pp. 11-18, 1993), incorporated by reference, has been described as powerful hybridization-based approach to genetic linkage mapping. GMS has applications both in the mapping of genetic traits and in the diagnosis and prevention of disease. What is currently impeding practical application of the GMS method is the lack of a sequence or map of the human (or animal model) genome that would provide densely spaced (e.g., 1 Mb) hybridization probes for the genome sampling step that scans the mismatched genome DNAs. Applicant's invention discloses a practical method for constructing such a sequence or map of a genome using the method described in the specification. In a preferred embodiment, densely spaced subsequences from the constructed sequence of a genome are used as hybridization probes in GMS. In an alternative embodiment, densely spaced clones (or subsequences therefrom) from the constructed map of a genome are used as hybridization probes in GMS.

1.4.11.10 Application: Reliable maps from unreliable data

A sequence and map of a genome is determined by the method described herein. It is generally believed that such maps can be reliably constructed only from highly reliable and relatively complete data. This belief adds considerably to the time, expense, and effort currently expended in constructing genome maps. However, the method described herein discloses a novel mechanism for constructing highly reliable maps from unreliable and incomplete data (J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," in *Automata*

Studies, C. E. Shannon and J. McCarthy, ed. Princeton, N.J.: Princeton University Press, 1956, pp. 43-98), incorporated by reference.

Specifically:

In step 6, table A's long-range characterization of the clone library can be comprised of very noisy, highly unreliable hybridization data exhibiting large error rates.

In step 9, table B's characterization of the long-range probe library can be sparsely sampled. In some embodiments, a. gtoreq. 1 Mb average inter-bin distance suffices for accurate mapping and contig construction.

In step 14, table D's short-range characterization of the clone library has a high tolerance for data errors.

This unobvious result is due to the considerable redundancy in the three data tables, and to the noise filtering and consistency cross-checking capabilities of the analysis methods:

In step 11, table C is a highly reliable binning because the clean PCR-based data table B is used as a global corrective for the noisy complex hybridization-based data table A. This has been empirically demonstrated for human chromosome 11.

In step 16, table E is a highly reliable contiging because every clone has been probed with both long-range and short-range data. Therefore, the global binning information relaxes the requirements on the short-range probings: useful comparisons can be made within a relatively small bin region using imperfect data.

1.4.11.11 Application: Mutation Detection

The techniques described herein will have a wide range of applications, particularly wherever desired to determine if a target nucleic acid has a particular nucleotide sequence or some other sequence differing from a known sequence. For example, one application of the inventions herein is found in mutation detection. These techniques may be applied in a wide variety of fields including diagnostics, forensics, bioanalytics, and others.

For example, assume a "wild-type" nucleic acid has the sequence 5'-N₁N₂N₃N₄ where, again, N refers to a monomer such as a nucleotide in a nucleic acid and the subscript refers to position number. Assume that a target nucleic acid is to be

evaluated to determine if it is the same as 5'-N₁N₂N₃N₄ or if it differs from this sequence, and so contains a mutation or mutant sequence. The target nucleic acid is initially exposed to an array of typically shorter probes, as discussed above.

Thereafter, one or more "core" sequences are identified, each of which would be expected to have a high binding affinity to the target, if the target does not contain a mutant sequence or mutation. In this particular example, one probe that would be expected to exhibit high binding affinity would be the complement to 5'-N₁N₂N₃ 3'-P₁P₂P₃, assuming a 3-mer array is utilized. Again, it will be recognized that the probes and/or the target may be part of a longer nucleic acid molecule.

As an initial screening tool, the absolute binding affinity of the target to the 3'-P₁P₂P₃ probe will be utilized to determine if the first three positions of the target are of the expected sequence. If the complement to 5'-N₁N₂N₃ does not exhibit strong binding to the target, it can be properly concluded that the target is not of the wild-type.

The single base mismatch profile can also be utilized according to the present invention to determine if the target contains a mutant or wild-type sequence. As shown herein, the single base mismatch plots for wild-type targets generally follow the typical, smile-shaped plot. Conversely, when the target has a mutation at a particular position, not only will the absolute binding affinity of the target to a particular core probe be less, but the single base mismatch characteristics will deviate from expected behavior.

According to one aspect of the invention, a substrate having a selected group of nucleic acids (otherwise referred to herein as a "library" of nucleic acids") is used in the determination of whether a particular nucleic acid is the same or different than a wild-type or other expected nucleic acid. Libraries of nucleic acids will normally be provided as an array of probes or "probe array." Such probe arrays are preferably formed on a single substrate in which the identity of a probe is determined by ways of its location on the substrate. Optionally, such substrates will not only determine if the nucleotide sequence of a target is the same as the wild-type, but it will also provide sequence information regarding the target. Such substrates will find use in fields noted above such as in forensics, diagnostics, and others. Merely by way of specific example, the invention may be utilized in diagnostics associated with sickle cell anemia detection, detection of any of the large number of P-53 mutations, for any of

the large number of cystic fibrosis mutations, for any particular variant sequence associated with the highly polymorphic HLA class 1 or class 2 genes (particularly class 2 DP, DQ and DR beta genes), as well as many other sequences associated with genetic diseases, genetic predisposition, and genetic evaluation.

When a substrate is to be used in such applications, it is not necessary to provide all of the possible nucleic acids of a particular length on the substrate. Instead, it will be necessary using the present invention to provide only a relatively small subset 45 of all the possible sequences. For example, suppose a target nucleic acid comprises a 5-base sequence of particular interest and that one wishes to develop a substrate that may be used to detect a single substitution in the 5-base sequence. According to one aspect of the invention, the substrate will be formed with the expected 5-base sequence formed on a surface thereof, along with all or most of the single base mismatch probes of the 5-base sequence. Accordingly, it will not be necessary to include all possible 5-base sequences on the substrate, although larger arrays will often be preferred. Typically, the length of the nucleic acid probes on the substrate according to the present invention will be between about 5 and 100 bases, between about 5 and 50 bases, between about 8 and 30 bases, or between about 8 and 15 bases.

By selection of the single base mismatch probes among all possible probes of a certain length, the number of probes on the substrate can be greatly limited. For example, in a 3-base sequence there are 69 possible DNA base sequences, but there will be only one exact complement to an expected sequence and 9 possible single base mismatch probes. By selecting only these probes, the diversity necessary for screening will be reduced. Preferably, but not necessarily, all of such single base mismatch probes are synthesized on a single substrate. While substrates will often be formed including other probes of interest in addition to the single base mismatches, such substrates will normally still have less than 50% of all the possible probes of n-bases, often less than 30% of all the possible probes of n-bases, often less than 20% of all the possible probes of n-bases, often less than 10% of the possible probes of n-bases, and often less than 5% of the possible probes of n-bases.

Nucleic acid probes will often be provided in a kit for analysis of a specific genetic sequence. According to one embodiment the kits will include a probe complementary to a target nucleic acid of interest. In addition, the kit will include

single base mismatches of the target. The kit will normally include one or more of C, G, T, A and/or U single base mismatches of such probe. Such kits will often be provided with appropriate instructions for use of the complementary probe and single base mismatches in determining the sequence of a particular nucleic acid sample in accordance with the teachings herein. According to one aspect of the invention, the kit provides for the complement to the target, along with only the single base mismatches. Such kits will often be utilized in assessing a particular sample of genetic material to determine if it indicates a particular genetic characteristic. For example, such kits may be utilized in the evaluation of a sample as mentioned above in the detection of sickle cell anemia, detection of any of the large number of P-53 mutations, detection of the large number of cystic fibrosis mutations, detection of particular variant sequence associated with the highly polymorphic HLA class 1 or class 2 genes (particularly class 2 DP, DQ and DR beta genes), as well as detection of many other sequences associated with genetic diseases, genetic predisposition, and genetic evaluation.

Accordingly, it is seen that substrates with probes selected according to the present invention will be capable of performing many mutation detection and other functions, but will need only a limited number of probes to perform such functions.

2. ANNOTATING

In one aspect this invention discloses the use of a relational database system for storing and manipulating biomolecular sequence information and storing and displaying genetic information, the database including genomic libraries for a plurality of types of organisms, the libraries having multiple genomic sequences, at least some of which represent open reading frames located along a contiguous sequence on each the plurality of organisms' genomes, and a user interface capable of receiving a selection of two or more of the genomic libraries for comparison and displaying the results of the comparison. Associated with the database is a software system that allows a user to determine the relative position of a selected gene sequence within a genome. The system allows execution of a method of displaying the genetic locus of a biomolecular sequence. The method involves providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The system also provides a user interface capable of receiving a selection of one or more probe open reading frames for use in determining homologous matches between such probe open reading frame(s) and the open reading frames in the genomic libraries, and displaying the results of the determination. An open reading frame for the sequence is selected and displayed together with adjacent open reading frames located upstream and downstream in the relative positions in which they occur on the contiguous sequence.

Also disclosed is a relational database system for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more protein function hierarchies. The hierarchies allow searches for sequences based upon a protein's biological function or molecular function. Also disclosed is a mechanism for automatically grouping new sequences into protein function hierarchies. This mechanism uses descriptive information obtained from "external hits" which are matches of stored sequences against gene sequences stored in an external database such as GenBank. The descriptive information provided with the external database is evaluated according to a specific algorithm and used to automatically group the external hits (or the sequences associated with the hits) in the categories. Ultimately, the biomolecular sequences stored in databases of this invention are provided with both descriptive

information from the external hit and category information from a relevant hierarchy or hierarchies.

Disclosed is a relational database system for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to association with one or more projects for obtaining full-length biomolecular sequences from shorter sequences. The relational database has sequence records containing information identifying one or more projects to which each of the sequence records belong. Each project groups together one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The computer system has a user interface allowing a user to selectively view information regarding one or more projects. The relational database also provides interfaces and methods for accessing and manipulating and analyzing project-based information.

Polymer sequences are assembled into bins. A first number of bins are populated with polymer sequences. The polymer sequences in each bin are assembled into one or more consensus sequences representative of the polymer sequences of the bin. The consensus sequences of the bins are compared to determine relationships, if any, between the consensus sequences of the bins. The bins are modified based on the relationships between the consensus sequences of the bins. The polymer sequences are reassembled in the modified bins to generate one or more modified consensus sequences for each bin representative of the modified bins. In another aspect of the invention, sequence similarities and dissimilarities are analyzed in a set of polymer sequences. Pairwise alignment data is generated for pairs of the polymer sequences. The pairwise alignment data defines regions of similarity between the pairs of polymer sequences with boundaries. Additional boundaries in particular polymer sequences are determined by applying at least one boundary from at least one pairwise alignment for one pair of polymer sequences to at least one other pairwise alignment for another pair of polymer sequences including one of the particular polymer sequences. Additional regions of similarity are generated based on the boundaries.

2.1. ANNOTATING - GENERAL METHODOLOGY

In one aspect this present invention relates generally to relational databases for storing and retrieving biological information. More particularly the invention relates to systems and methods for providing sequences of biological molecules in a relational format allowing retrieval in a client-server environment and for providing full-length cDNA sequences in a relational format allowing retrieval in a client-server environment.

Informatics is the study and application of computer and statistical techniques to the management of information. In genome projects, bioinformatics includes the development of methods to search databases quickly, to analyze nucleic acid sequence information, and to predict protein sequence, structure and function from DNA sequence data.

Increasingly, molecular biology is shifting from the laboratory bench to the computer desktop. Today's researchers require advanced quantitative analyses, database comparisons, and computational algorithms to explore the relationships between sequence and phenotype. Thus, by all accounts, researchers can not and will not be able to avoid using computer resources to explore gene expression, gene sequencing and molecular structure.

One use of bioinformatics involves studying an organism's genome to determine the sequence and placement of its genes and their relationship to other sequences and genes within the genome or to genes in other organisms. Another use of bioinformatics involves studying genes differentially or commonly expressed in different tissues or cell lines (e.g. normal and cancerous tissue).

Such information is of significant interest in biomedical and pharmaceutical research, for instance to assist in the evaluation of drug efficacy and resistance.

The sequence tag method involves generation of a large number (e.g., thousands) of Expressed Sequence Tags ("ESTs") from cDNA libraries (each produced from a different tissue or sample). ESTs are partial transcript sequences that may cover different parts of the cDNA(s) of a gene, depending on cloning and sequencing strategy. Each EST includes about 50 to 300 nucleotides. If it is assumed that the number of tags is proportional to the abundance of transcripts in the tissue or cell type used to make the cDNA library, then any variation in the relative frequency of those tags, stored in computer databases, can be used to detect the differential abundance and potentially the expression of the corresponding genes.

To make genomic and EST information manipulation easy to perform and understand, sophisticated computer database systems have been developed. In one database system, developed by Incyte Pharmaceuticals, Inc. of Palo Alto, CA, genomic sequence data and the abundance levels of mRNA species represented in a given sample is electronically recorded and annotated with information available from public sequence databases such as GenBank. Examples of such databases include GenBank (NCBI) and TIGR. The resulting information is stored in a relational database that may be employed to determine relationships between sequences and genes within and among genomes and establish a cDNA profile for a given tissue and to evaluate changes in gene expression caused by disease progression, pharmacological treatment, aging, etc.

In one database system, developed by Incyte Pharmaceuticals, Inc. of Palo Alto, Calif., abundance levels of mRNA species represented in a given sample are electronically recorded and annotated with information available from public sequence databases such as GenBank. The resulting information is stored in a relational database that may be employed to establish a cDNA profile for a given tissue and to evaluate changes in gene expression caused by disease progression, pharmacological treatment, aging, etc.

Genetic information for a number of organisms has been catalogued in computer databases. Genetic databases for organisms such as *Eschericia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, and *Mycoplasma pneumoniae*, among others, are publicly available. At present, however, complete sequence data is available for relatively few species, and the ability to manipulate sequence data within and between species and databases is limited.

While genetic data processing and relational database systems such as those developed by Incyte Pharmaceuticals, Inc. provide great power and flexibility in analyzing genetic information and gene expression information, this area of technology is still in its infancy and further improvements in genetic data processing and relational database systems and their content will help accelerate biological research for numerous applications.

In genome projects, bioinformatics includes the development of methods to search databases quickly, to analyze nucleic acid sequence information, and to predict protein sequence and structure from DNA sequence data. Increasingly, molecular biology is shifting from the laboratory bench to the computer desktop. Advanced quantitative

analyses, database comparisons, and computational algorithms are needed to explore the relationships between sequence and phenotype.

2.2. ANNOTATING – EXEMPLARY ASPECTS

The annotation methods of this invention include those described in PCT patent publication Nos. 98/26407, 98/26408, and 99/49403 and United States Patent Nos. 6,023,659 and 5,953,727 and are herein incorporated by reference in their entirety to the same extent as if each individual patent or patent application were specifically and individually indicated to be incorporated by reference in its entirety.

Thus, in one aspect, this present invention provides relational database systems for storing and analyzing biomolecular sequence information together with biological annotations detailing the source and interpretation the sequence data. The present invention provides a powerful database tool for drug development and other research and development purposes.

The present invention provides relational database systems for storing and analyzing biomolecular sequence information together with biological detailing the source and interpretation the sequence data. Disclosed is a relational database systems for storing and displaying genetic information.

Associated with the database is a software system the allows a user to determine the relative position of a selected gene sequence within a genome. The system allows execution of a method of displaying the genetic locus of a biomolecular sequence. The method involves providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. An open reading frame for the sequence is selected and displayed together with adjacent open reading frames located upstream and downstream in the relative positions in which they occur on the contiguous sequence.

The invention provides a method of displaying the genetic locus of a biomolecular sequence. The method involve providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The method further involves identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame.

The adjacent open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence, textually and/or graphically. The method of the invention may be practiced with sequences from microbial organisms, and the sequences may include nucleic acid or protein sequences.

The invention also provides a computer system including a database having multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome.

The computer system also includes a user interface capable of identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame. The adjacent open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence. The user interface may also be capable of detecting a scrolling command, and based upon the direction and magnitude of the scrolling command, identifying a new selected open reading frame from the contiguous sequence.

The invention further provides a computer program product comprising a computer-usable medium having computer-readable program code embodied thereon relating to a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The computer program product includes computer-readable program code for identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame. The adjacent open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence.

Comparative Genomics is a feature of the database system of the present invention which allows a user to compare the sequence data of sets of different organism types. Comparative searches may be formulated in a number of ways using the Comparative Genomics feature. For example, genes common to a set of organisms may be identified through a "commonality" query, and genes unique to one of a set of organisms may be identified through a "subtraction" query.

Electronic Southern is a feature of the present database system which is useful for identifying genomic libraries in which a given gene or ORF exists.

A Southern analysis is a conventional molecular biology technique in which a nucleic acid of known sequence is used to identify matching (complementary) sequences in a sample of nucleic acid to be analyzed. Like their laboratory counterparts, Electronic Southern according to the present invention may be used to locate homologous matches between a "probe" DNA sequence and a large number of DNA sequences in one or more libraries.

The present invention provides a method of comparing genetic complements of different types of organisms. The method involves providing a database having sequence libraries with multiple biomolecular sequences for different types of organisms, where at least some of the sequences represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of two or more of the sequence libraries for comparison, determining open reading frames common or unique to the selected sequence libraries, and displaying the results of the determination.

The invention also provides a method of comparing genomic complements of different types of organisms. The method involves providing a database having genomic sequence libraries with multiple biomolecular sequences for different types of organisms, where at least some of the sequences represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of two or more of the sequence libraries for comparison, determining sequences common or unique to the selected sequence libraries, and displaying the results of the determination.

The invention further provides a computer system including a database containing genomic libraries for different types of organisms, which libraries have multiple genomic sequences, at least some of which representing open reading frames located along one or more contiguous sequences on each the organisms' genomes. The system also includes a user interface capable of receiving a selection of two or more genomic libraries for comparison and displaying the results of the comparison.

Another aspect of the present invention provides a method of identifying libraries in which a given gene exists. The method involves providing a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic

sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The method further involves receiving a selection of one or more probe sequences, determining homologous matches between the selected probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

The invention also provides a computer system including a database including genomic libraries for one or more types of organisms, which libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The system also includes a user interface capable of receiving a selection of one or more probe sequences for use in determining homologous matches between one or more probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

Also provided is a computer program product including a computer-usable medium having computer-readable program code embodied thereon relating to a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The computer program product includes computer-readable program code for providing, within a computing system, an interface for receiving a selection of two or more genomic libraries for comparison, determining sequences common or unique to the selected genomic libraries, and displaying the results of the determination.

Additionally provided is a computer program product including a computer-usable medium having computer-readable program code embodied thereon relating to a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The computer program product includes computer-readable program code for providing, within a computing system, an interface for receiving a selection of one or more probe open reading frames, determining homologous matches between the probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

The invention further provides a method of presenting the genetic complement of an organism. The method involves providing a database including sequence libraries for a plurality of types of organisms, where the libraries have multiple biomolecular sequences,

at least some of which represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of one of the sequence libraries, determining open reading frames within the selected sequence library, and displaying the results as one or more unique identifiers for groups of related opening reading frames.

The present invention provides relational database systems for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more protein function hierarchies. The hierarchies are provided to allow carefully tailored searches for sequences based upon a protein's biological function or molecular function. To make this capability available in large sequence databases, the invention provides a mechanism for automatically grouping new sequences into protein function hierarchies. This mechanism takes advantage of descriptive information obtained from "external hits" which are matches of stored sequences against gene sequences stored in an external database such as GenBank. The descriptive information provided with GenBank is evaluated according to a specific algorithm and used to automatically group the external hits (or the sequences associated with the hits) in the categories. Ultimately, the biomolecular sequences stored in databases of this invention are provided with both descriptive information from the external hit and category information from a relevant hierarchy or hierarchies.

The invention provides a computer system having a database containing records pertaining to a plurality of biomolecular sequences. At least some of the biomolecular sequences are grouped into a first hierarchy of protein function categories, the protein function categories specifying biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy. The hierarchy includes a first set of protein function categories specifying biological functions at a cellular level, and a second set of protein function categories specifying biological functions at a level above the cellular level. The computer system of the invention also includes a user interface allowing a user to selectively view information regarding the plurality of biomolecular sequences as it relates to the first hierarchy. The computer system may also include additional protein function categories based, for example, on molecular or enzymatic function of proteins. The biomolecular sequences may include nucleic acid or amino acid sequences. Some of said biomolecular sequences may be provided as part of one or more projects for obtaining

full-length gene sequences from shorter sequences, and the database records may contain information about such projects.

The invention also provides a method of using a computer system to present information pertaining to a plurality of biomolecular sequence records stored in a database. The method involves displaying a list of the records or a field for entering information identifying one or more of the records, identifying one or more of the records that a user has selected from the list or field, matching the one or more selected records with one or more protein function categories from a first hierarchy of protein function categories into which at least some of the biomolecular sequence records are grouped, and displaying the one or more categories matching the one or more selected records. The protein function categories specify biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy includes a first set of protein function categories specifying biological functions at a cellular level, and a second set of protein function categories specifying biological functions at a tissue level. The method may also involve matching the records against other protein function hierarchies, such as hierarchies based on molecular and/or enzymatic function, and displaying the results. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Additionally, the invention provides a method of using a computer system to present information pertaining to a plurality of biomolecular sequence records stored in a database. The method involves displaying a list of one or more protein biological function categories from a first hierarchy of protein biological function categories into which at least some of the biomolecular sequence records are grouped, identifying one or more of the protein biological function categories that a user has selected from the list, matching the one or more selected protein biological function categories with one or more biomolecular sequence records which are grouped in the selected protein biological function categories, and displaying the one or more sequence records matching the one or more selected protein biological function categories. The protein biological function categories specify biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy includes a first set of protein biological function categories specifying biological functions at a cellular level, and a second set of protein biological function categories specifying biological functions at a tissue level. The method

may also involve matching the records against other protein function hierarchies, such as hierarchies based on molecular and/or enzymatic function, and displaying the results. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Another aspect of the invention provides a database system having a plurality of internal records. The database includes a plurality of sequence records specifying biomolecular sequences, at least some of which records reference hits to an external database, which hits specify genes having sequences that at least partially match those of the biomolecular sequences. The database also includes a plurality of external hit records specifying the hits to the external database, and at least some of the records reference protein function hierarchy categories which specify at least one of biological functions of proteins or molecular functions of proteins. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Further aspects of the present invention provide a method of using a computer system and a computer readable medium having program instructions to automatically categorize biomolecular sequence records into protein function categories in an internal database. The method and program involve receiving descriptive information about a biomolecular sequence in the internal database from a record in an external database pertaining to a gene having a sequence that at least partially matches that of the biomolecular sequence. Next, a determination is made whether the descriptive information contains one or more terms matching one or more keywords associated with a first protein function category, the keywords being terms consistent with a classification in the first protein function category. When at least one keyword is found to match a term in the descriptive information, a determination is made whether the descriptive information contains a term matching one or more anti-keywords associated with the first protein function category, the anti-keywords being terms inconsistent with a classification in the first protein function category. Then, the biomolecular sequence is grouped in the first protein function category when the descriptive information contains a term matching a keyword but contains no term matching an anti-keyword.

with reference to the drawings,

The present invention provides relational database systems for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more characteristics. The sequence information of the database is generated by one or more "projects" which are concerned with identifying the full-length coding sequence of a gene (i.e., mRNA). The projects involve the extension of an initial sequenced portion of a clone of a gene of interest (e.g., an EST) by a variety of methods which use conventional molecular biological techniques, recently developed adaptations of these techniques, and certain novel database applications. Data accumulated in these projects may be provided to the database of the present invention throughout the course of the projects and may be available to database users (subscribers) throughout the course of these projects for research, product (i.e., drug) development, and other purposes.

In a preferred embodiment, the database of the present invention and its associated projects may provide sequence and related data in amounts and forms not previously available. The present invention preferably makes partial and full-length sequence information for a given gene available to a user both during the course of the data acquisition and once the full-length sequence of the gene has been elucidated. The database also preferably provides a variety of tools for analysis and manipulation of the data, including Northern analysis and Expression summaries. The present invention should permit more complete and accurate annotation of sequence data, as well as the study of relationships between genes of different tissues, systems or organisms, and ultimately detailed expression studies of full-length gene sequences.

The invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong. Each project groups together one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The computer system also has a user interface allowing a user to selectively view information regarding one or more projects. The biomolecular sequences may include nucleic acid or amino acid sequences. The user interface may allow users to view at least three levels of project information including a project information results level listing at least some of the projects in said database, a sequence information results level listing at least some of the sequences associated with a given project, and a sequence retrieval results level sequentially listing monomers which comprise a given sequence.

A method of using a computer system and a computer program product to present information pertaining to a plurality of sequence records stored in a database are also provided by the present invention. The sequence records contain information identifying one or more projects to which each of the sequence records belong. Each of the projects groups one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method and program involve providing an interface for entering query information relating to one or more projects, locating data corresponding to the entered query information, and displaying the data corresponding to the entered query information.

Additionally, the invention provides a method of using a computer system to present information pertaining to a plurality of sequence records stored in a database. The sequence records contains information identifying one or more projects to which each of the sequence records belong. Each of the projects groups one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method involves displaying a list of one or more project identifiers, determining which project identifier or identifiers from the list is selected by a user, then displaying a second list of one or more biomolecular sequence identifiers associated with the selected project identifier or identifiers, determining which sequence identifier or identifiers from the second list has been selected by a user, and displaying a third list of one or more sequences corresponding to the selected sequence identifier or identifiers. Following the display of the third list, a determination may be made whether and which sequence from the third list has been selected by a user. If a sequence is selected, a sequence alignment search of the selected sequence against other databased sequences may be initiated, and the results of the alignment search displayed.

For Electronic Northern analysis, the invention further provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of said projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The system also has a user interface capable of allowing a user to select one or more project identifiers or project member identifiers specifying one or more sequences to be compared with one or more cDNA sequence libraries, and displaying matches resulting from that comparison.

A method of using a computer system to present comparative information pertaining to a plurality of sequence records stored in a database is also provided by the present invention. The sequence records contain information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method involves providing an interface capable of allowing a user to select one or more project identifiers or project member identifiers specifying one or more sequences, comparing the one or more specified sequences with one or more cDNA sequence libraries, and displaying matches resulting from the comparison.

In addition, for Expression analysis, the invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The system also has a user interface allowing a user to view expression information pertaining to the projects by selecting one or more expression categories for a query, and displaying the result of the query.

A method of using a computer system to view expression information pertaining to one or more projects, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence, is also provided in accordance with the present invention. The computer system includes a database storing a plurality of sequence records, the sequence records containing information identifying one or more projects to which each of the sequence records belong. The method involves providing an interface which allows a user to select one or more expression categories as a query, locating projects belonging to the selected one or more expression categories, and displaying a list of located projects.

Finally, the present invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. This computer system has a user interface allowing a user to selectively view information regarding said one or more projects and which displays information to a user in a format common to one or more other sequence databases.

These and other features and advantages of the invention will be described in more detail below with reference to the drawings.

Polymer sequences are assembled into bins. A first number of bins are populated with polymer sequences. The polymer sequences in each bin are assembled into one or more consensus sequences representative of the polymer sequences of the bin. The consensus sequences of the bins are compared to determine relationships, if any, between the consensus sequences. The bins are modified based on the relationships between the consensus sequences. The polymer sequences are reassembled in the modified bins to generate one or more modified consensus sequences for each bin representative of the modified bins.

In another aspect of the invention, sequence similarities and dissimilarities are analyzed in a set of polymer sequences. Pairwise alignment data is generated for pairs of the polymer sequences. The pairwise alignment data defines regions of similarity between the pairs of polymer sequences with boundaries. Additional boundaries in particular polymer sequences are determined by applying at least one boundary from at least one pairwise alignment for one pair of polymer sequences to at least one other pairwise alignment for another pair of polymer sequences including one of the particular polymer sequences. Additional regions of similarity are generated based on the boundaries

2.3. ANNOTATING - PREFERRED EMBODIMENTS

Generally, the present invention provides an improved relational database for storing and manipulating genomic sequence information. While the invention is described in terms of a database optimized for microbial data, it is by no means so limited. The invention may be employed to investigate data from various sources. For example, the invention covers databases optimized for other sources of sequence data, such as animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.), plant sequences and microbial sequences. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, that the present invention may be practiced without limitation to some of the specific details presented herein.

Generally, the present invention provides an improved relational database for storing sequence information. The invention may be employed to investigate data from various sources. For example, it may catalogue animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.), plant sequences, and microbial sequences.

3. DIRECTED EVOLUTION METHODS

In one aspect the invention described herein is directed to the use of repeated cycles of reductive reassortment, recombination and selection which allow for the directed molecular evolution of highly complex linear sequences, such as DNA, RNA or proteins thorough recombination.

In vivo shuffling of molecules can be performed utilizing the natural property of cells to recombine multimers. While recombination *in vivo* has provided the major natural route to molecular diversity, genetic recombination remains a relatively complex process that involves 1) the recognition of homologies; 2) strand cleavage, strand invasion, and metabolic steps leading to the production of recombinant chiasma; and finally 3) the resolution of chiasma into discrete recombined molecules. The formation of the chiasma requires the recognition of homologous sequences.

In a preferred embodiment, the invention relates to a method for producing a hybrid polynucleotide from at least a first polynucleotide and a second polynucleotide. The present invention can be used to produce a hybrid polynucleotide by introducing at least a first polynucleotide and a second polynucleotide which share at least one region of partial sequence homology into a suitable host cell. The regions of partial sequence homology promote processes which result in sequence reorganization producing a hybrid polynucleotide. The term "hybrid polynucleotide", as used herein, is any nucleotide sequence which results from the method of the present invention and contains sequence from at least two original polynucleotide sequences. Such hybrid polynucleotides can result from intermolecular recombination events which promote sequence integration between DNA molecules. In addition, such hybrid polynucleotides can result from intramolecular reductive reassortment processes which utilize repeated sequences to alter a nucleotide sequence within a DNA molecule.

The invention provides a means for generating hybrid polynucleotides which may encode biologically active hybrid polypeptides. In one aspect, the original polynucleotides encode biologically active polypeptides. The method of the invention produces new hybrid polypeptides by utilizing cellular processes which integrate the sequence of the

original polynucleotides such that the resulting hybrid polynucleotide encodes a polypeptide demonstrating activities derived from the original biologically active polypeptides. For example, the original polynucleotides may encode a particular enzyme from different microorganisms. An enzyme encoded by a first polynucleotide from one organism may, for example, function effectively under a particular environmental condition, e.g. high salinity. An enzyme encoded by a second polynucleotide from a different organism may function effectively under a different environmental condition, such as extremely high temperatures. A hybrid polynucleotide containing sequences from the first and second original polynucleotides may encode an enzyme which exhibits characteristics of both enzymes encoded by the original polynucleotides. Thus, the enzyme encoded by the hybrid polynucleotide may function effectively under environmental conditions shared by each of the enzymes encoded by the first and second polynucleotides, e.g., high salinity and extreme temperatures.

Enzymes encoded by the original polynucleotides of the invention include, but are not limited to; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. A hybrid polypeptide resulting from the method of the invention may exhibit specialized enzyme activity not displayed in the original enzymes. For example, following recombination and/or reductive reassortment of polynucleotides encoding hydrolase activities, the resulting hybrid polypeptide encoded by a hybrid polynucleotide can be screened for specialized hydrolase activities obtained from each of the original enzymes, i.e. the type of bond on which the hydrolase acts and the temperature at which the hydrolase functions. Thus, for example, the hydrolase may be screened to ascertain those chemical functionalities which distinguish the hybrid hydrolase from the original hydrolyases, such as: (a) amide (peptide bonds), i.e. proteases; (b) ester bonds, i.e. esterases and lipases; (c) acetals, i.e., glycosidases and, for example, the temperature, pH or salt concentration at which the hybrid polypeptide functions.

Sources of the original polynucleotides may be isolated from individual organisms ("isolates"), collections of organisms that have been grown in defined media ("enrichment cultures"), or, most preferably, uncultivated organisms ("environmental samples"). The use of a culture-independent approach to derive polynucleotides encoding novel

bioactivities from environmental samples is most preferable since it allows one to access untapped resources of biodiversity.

"Environmental libraries" are generated from environmental samples and represent the collective genomes of naturally occurring organisms archived in cloning vectors that can be propagated in suitable prokaryotic hosts. Because the cloned DNA is initially extracted directly from environmental samples, the libraries are not limited to the small fraction of prokaryotes that can be grown in pure culture. Additionally, a normalization of the environmental DNA present in these samples could allow more equal representation of the DNA from all of the species present in the original sample. This can dramatically increase the efficiency of finding interesting genes from minor constituents of the sample which may be under-represented by several orders of magnitude compared to the dominant species.

For example, gene libraries generated from one or more uncultivated microorganisms are screened for an activity of interest. Potential pathways encoding bioactive molecules of interest are first captured in prokaryotic cells in the form of gene expression libraries. Polynucleotides encoding activities of interest are isolated from such libraries and introduced into a host cell. The host cell is grown under conditions which promote recombination and/or reductive reassortment creating potentially active biomolecules with novel or enhanced activities.

The microorganisms from which the polynucleotide may be prepared include prokaryotic microorganisms, such as Eubacteria and Archaeobacteria, and lower eukaryotic microorganisms such as fungi, some algae and protozoa. Polynucleotides may be isolated from environmental samples in which case the nucleic acid may be recovered without culturing of an organism or recovered from one or more cultured organisms. In one aspect, such microorganisms may be extremophiles, such as hyperthermophiles, psychrophiles, psychrotrophs, halophiles, barophiles and acidophiles. Polynucleotides encoding enzymes isolated from extremophilic microorganisms are particularly preferred. Such enzymes may function at temperatures above 100°C in terrestrial hot springs and deep sea thermal vents, at temperatures below 0°C in arctic waters, in the saturated salt environment of the Dead Sea, at pH values around 0 in coal deposits and geothermal sulfur-rich springs, or at pH

values greater than 11 in sewage sludge. For example, several esterases and lipases cloned and expressed from extremophilic organisms show high activity throughout a wide range of temperatures and pHs.

Polynucleotides selected and isolated as hereinabove described are introduced into a suitable host cell. A suitable host cell is any cell which is capable of promoting recombination and/or reductive reassortment. The selected polynucleotides are preferably already in a vector which includes appropriate control sequences. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or preferably, the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis et al, 1986).

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as *E. coli*, *Streptomyces*, *Salmonella typhimurium*; fungal cells, such as yeast; insect cells such as *Drosophila S2* and *Spodoptera Sf9*; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; and plant cells. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

With particular references to various mammalian cell culture systems that can be employed to express recombinant protein, examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described in "SV40-transformed simian cells support the replication of early SV40 mutants" (Gluzman, 1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Host cells containing the polynucleotides of interest can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting

transformants or amplifying genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan. The clones which are identified as having the specified enzyme activity may then be sequenced to identify the polynucleotide sequence encoding an enzyme having the enhanced activity.

In another aspect, it is envisioned the method of the present invention can be used to generate novel polynucleotides encoding biochemical pathways from one or more operons or gene clusters or portions thereof. For example, bacteria and many eukaryotes have a coordinated mechanism for regulating genes whose products are involved in related processes. The genes are clustered, in structures referred to as "gene clusters," on a single chromosome and are transcribed together under the control of a single regulatory sequence, including a single promoter which initiates transcription of the entire cluster. Thus, a gene cluster is a group of adjacent genes that are either identical or related, usually as to their function. An example of a biochemical pathway encoded by gene clusters are polyketides. Polyketides are molecules which are an extremely rich source of bioactivities, including antibiotics (such as tetracyclines and erythromycin), anti-cancer agents (daunomycin), immunosuppressants (FK506 and rapamycin), and veterinary products (monensin). Many polyketides (produced by polyketide synthases) are valuable as therapeutic agents. Polyketide synthases are multifunctional enzymes that catalyze the biosynthesis of an enormous variety of carbon chains differing in length and patterns of functionality and cyclization. Polyketide synthase genes fall into gene clusters and at least one type (designated type I) of polyketide synthases have large size genes and enzymes, complicating genetic manipulation and *in vitro* studies of these genes/proteins.

The ability to select and combine desired components from a library of polyketides, or fragments thereof, and postpolyketide biosynthesis genes for generation of novel polyketides for study is appealing. The method of the present invention makes it possible to facilitate the production of novel polyketide synthases through intermolecular recombination.

Preferably, gene cluster DNA can be isolated from different organisms and ligated into vectors, particularly vectors containing expression regulatory sequences which can

control and regulate the production of a detectable protein or protein-related array activity from the ligated gene clusters. Use of vectors which have an exceptionally large capacity for exogenous DNA introduction are particularly appropriate for use with such gene clusters and are described by way of example herein to include the f-factor (or fertility factor) of *E. coli*. This f-factor of *E. coli* is a plasmid which affect high-frequency transfer of itself during conjugation and is ideal to achieve and stably propagate large DNA fragments, such as gene clusters from mixed microbial samples. Once ligated into an appropriate vector, two or more vectors containing different polyketide synthase gene clusters can be introduced into a suitable host cell. Regions of partial sequence homology shared by the gene clusters will promote processes which result in sequence reorganization resulting in a hybrid gene cluster. The novel hybrid gene cluster can then be screened for enhanced activities not found in the original gene clusters.

Therefore, in a preferred embodiment, the present invention relates to a method for producing a biologically active hybrid polypeptide and screening such a polypeptide for enhanced activity by:

- 1) introducing at least a first polynucleotide in operable linkage and a second polynucleotide in operable linkage, said at least first polynucleotide and second polynucleotide sharing at least one region of partial sequence homology, into a suitable host cell;
- 2) growing the host cell under conditions which promote sequence reorganization resulting in a hybrid polynucleotide in operable linkage;
- 3) expressing a hybrid polypeptide encoded by the hybrid polynucleotide;
- 4) screening the hybrid polypeptide under conditions which promote identification of enhanced biological activity; and
- 5) isolating the a polynucleotide encoding the hybrid polypeptide.

Methods for screening for various enzyme activities are known to those of skill in the art and discussed throughout the present specification. Such methods may be employed when isolating the polypeptides and polynucleotides of the present invention.

As representative examples of expression vectors which may be used there may be mentioned viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids,

bacterial artificial chromosomes, viral DNA (e.g. vaccinia, adenovirus, fowl pox virus, pseudorabies and derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, aspergillus and yeast). Thus, for example, the DNA may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used as long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed with the present invention.

A preferred type of vector for use in the present invention contains an f-factor origin replication. The f-factor (or fertility factor) in *E. coli* is a plasmid which effects high frequency transfer of itself during conjugation and less frequent transfer of the bacterial chromosome itself. A particularly preferred embodiment is to use cloning vectors, referred to as "fosmids" or bacterial artificial chromosome (BAC) vectors. These are derived from *E. coli* f-factor which is able to stably integrate large segments of genomic DNA. When integrated with DNA from a mixed uncultured environmental sample, this makes it possible to achieve large genomic fragments in the form of a stable "environmental DNA library."

Another preferred type of vector for use in the present invention is a cosmid vector. Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. Cloning into cosmid vectors is described in detail in "Molecular Cloning: A laboratory Manual" (Sambrook et al, 1989).

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct RNA synthesis. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40,

LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers.

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in *E. coli*.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), -f actor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The cloning strategy permits expression via both vector driven and endogenous promoters; vector promotion may be important with expression of genes whose endogenous promoter will not function in *E. coli*.

The DNA isolated or derived from microorganisms can preferably be inserted into a vector or a plasmid prior to probing for selected DNA. Such vectors or plasmids are preferably those containing expression regulatory sequences, including promoters, enhancers and the like. Such polynucleotides can be part of a vector and/or a composition and still be isolated, in that such vector or composition is not part of its natural

environment. Particularly preferred phage or plasmid and methods for introduction and packaging into them are described in detail in the protocol set forth herein.

The selection of the cloning vector depends upon the approach taken, for example, the vector can be any cloning vector with an adequate capacity to multiply repeated copies of a sequence, or multiple sequences that can be successfully transformed and selected in a host cell. One example of such a vector is described in "Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts" (Alting-Mecs and Short, 1993). Propagation/maintenance can be by an antibiotic resistance carried by the cloning vector. After a period of growth, the naturally abbreviated molecules are recovered and identified by size fractionation on a gel or column, or amplified directly. The cloning vector utilized may contain a selectable gene that is disrupted by the insertion of the lengthy construct. As reductive reassortment progresses, the number of repeated units is reduced and the interrupted gene is again expressed and hence selection for the processed construct can be applied. The vector may be an expression/selection vector which will allow for the selection of an expressed product possessing desirable biologically properties. The insert may be positioned downstream of a functional promotor and the desirable property screened by appropriate means.

In vivo reassortment is focused on "inter-molecular" processes collectively referred to as "recombination" which in bacteria, is generally viewed as a "RecA-dependent" phenomenon. The present invention can rely on recombination processes of a host cell to recombine and re-assort sequences, or the cells' ability to mediate reductive processes to decrease the complexity of quasi-repeated sequences in the cell by deletion. This process of "reductive reassortment" occurs by an "intra-molecular", RecA-independent process.

Therefore, in another aspect of the present invention, novel polynucleotides can be generated by the process of reductive reassortment. The method involves the generation of constructs containing consecutive sequences (original encoding sequences), their insertion into an appropriate vector, and their subsequent introduction into an appropriate host cell. The reassortment of the individual molecular identities occurs by combinatorial processes between the consecutive sequences in the construct possessing regions of

homology, or between quasi-repeated units. The reassortment process recombines and/or reduces the complexity and extent of the repeated sequences, and results in the production of novel molecular species. Various treatments may be applied to enhance the rate of reassortment. These could include treatment with ultra-violet light, or DNA damaging chemicals, and/or the use of host cell lines displaying enhanced levels of "genetic instability". Thus the reassortment process may involve homologous recombination or the natural property of quasi-repeated sequences to direct their own evolution.

Repeated or "quasi-repeated" sequences play a role in genetic instability. In the present invention, "quasi-repeats" are repeats that are not restricted to their original unit structure. Quasi-repeated units can be presented as an array of sequences in a construct; consecutive units of similar sequences. Once ligated, the junctions between the consecutive sequences become essentially invisible and the quasi-repetitive nature of the resulting construct is now continuous at the molecular level. The deletion process the cell performs to reduce the complexity of the resulting construct operates between the quasi-repeated sequences. The quasi-repeated units provide a practically limitless repertoire of templates upon which slippage events can occur. The constructs containing the quasi-repeats thus effectively provide sufficient molecular elasticity that deletion (and potentially insertion) events can occur virtually anywhere within the quasi-repetitive units.

When the quasi-repeated sequences are all ligated in the same orientation, for instance head to tail or vice versa, the cell cannot distinguish individual units. Consequently, the reductive process can occur throughout the sequences. In contrast, when for example, the units are presented head to head, rather than head to tail, the inversion delineates the endpoints of the adjacent unit so that deletion formation will favor the loss of discrete units. Thus, it is preferable with the present method that the sequences are in the same orientation. Random orientation of quasi-repeated sequences will result in the loss of reassortment efficiency, while consistent orientation of the sequences will offer the highest efficiency. However, while having fewer of the contiguous sequences in the same orientation decreases the efficiency, it may still provide sufficient elasticity for the effective recovery of novel molecules. Constructs can be made with the quasi-repeated sequences in the same orientation to allow higher efficiency.

Sequences can be assembled in a head to tail orientation using any of a variety of methods, including the following:

- a) Primers that include a poly-A head and poly-T tail which when made single-stranded would provide orientation can be utilized. This is accomplished by having the first few bases of the primers made from RNA and hence easily removed RNaseH.
- b) Primers that include unique restriction cleavage sites can be utilized. Multiple sites, a battery of unique sequences, and repeated synthesis and ligation steps would be required.
- c) The inner few bases of the primer could be thiolated and an exonuclease used to produce properly tailed molecules.

The recovery of the re-assorted sequences relies on the identification of cloning vectors with a reduced RI. The re-assorted encoding sequences can then be recovered by amplification. The products are re-cloned and expressed. The recovery of cloning vectors with reduced RI can be effected by:

- 1) The use of vectors only stably maintained when the construct is reduced in complexity.
- 2) The physical recovery of shortened vectors by physical procedures. In this case, the cloning vector would be recovered using standard plasmid isolation procedures and size fractionated on either an agarose gel, or column with a low molecular weight cut off utilizing standard procedures.
- 3) The recovery of vectors containing interrupted genes which can be selected when insert size decreases.
- 4) The use of direct selection techniques with an expression vector and the appropriate selection.

Encoding sequences (for example, genes) from related organisms may demonstrate a high degree of homology and encode quite diverse protein products. These types of sequences are particularly useful in the present invention as quasi-repeats. However, while the examples illustrated below demonstrate the reassortment of nearly identical original encoding sequences (quasi-repeats), this process is not limited to such nearly identical repeats.

The following example demonstrates the method of the invention. Encoding nucleic acid sequences (quasi-repeats) derived from three (3) unique species are depicted. Each sequence encodes a protein with a distinct set of properties. Each of the sequences differs by a single or a few base pairs at a unique position in the sequence which are designated "A", "B" and "C". The quasi-repeated sequences are separately or collectively amplified and ligated into random assemblies such that all possible permutations and combinations are available in the population of ligated molecules. The number of quasi-repeat units can be controlled by the assembly conditions. The average number of quasi-repeated units in a construct is defined as the repetitive index (RI).

Once formed, the constructs may, or may not be size fractionated on an agarose gel according to published protocols, inserted into a cloning vector, and transfected into an appropriate host cell. The cells are then propagated and "reductive reassortment" is effected. The rate of the reductive reassortment process may be stimulated by the introduction of DNA damage if desired. Whether the reduction in RI is mediated by deletion formation between repeated sequences by an "intra-molecular" mechanism, or mediated by recombination-like events through "inter-molecular" mechanisms is immaterial. The end result is a reassortment of the molecules into all possible combinations.

Optionally, the method comprises the additional step of screening the library members of the shuffled pool to identify individual shuffled library members having the ability to bind or otherwise interact (e.g., such as catalytic antibodies) with a predetermined macromolecule, such as for example a proteinaceous receptor, peptide oligosaccharide, viron, or other predetermined compound or structure.

The displayed polypeptides, antibodies, peptidomimetic antibodies, and variable region sequences that are identified from such libraries can be used for therapeutic, diagnostic, research and related purposes (e.g., catalysts, solutes for increasing osmolarity of an aqueous solution, and the like), and/or can be subjected to one or more additional cycles of shuffling and/or affinity selection. The method can be modified such that the step of selecting for a phenotypic characteristic can be other than of binding affinity for a

predetermined molecule (e.g., for catalytic activity, stability oxidation resistance, drug resistance, or detectable phenotype conferred upon a host cell).

The present invention provides a method for generating libraries of displayed antibodies suitable for affinity interactions screening. The method comprises (1) obtaining first a plurality of selected library members comprising a displayed antibody and an associated polynucleotide encoding said displayed antibody, and obtaining said associated polynucleotide encoding for said displayed antibody and obtaining said associated polynucleotides or copies thereof, wherein said associated polynucleotides comprise a region of substantially identical variable region framework sequence, and (2) introducing said polynucleotides into a suitable host cell and growing the cells under conditions which promote recombination and reductive reassortment resulting in shuffled polynucleotides. CDR combinations comprised by the shuffled pool are not present in the first plurality of selected library members, said shuffled pool composing a library of displayed antibodies comprising CDR permutations and suitable for affinity interaction screening. Optionally, the shuffled pool is subjected to affinity screening to select shuffled library members which bind to a predetermined epitope (antigen) and thereby selecting a plurality of selected shuffled library members. Further, the plurality of selectively shuffled library members can be shuffled and screened iteratively, from 1 to about 1000 cycles or as desired until library members having a desired binding affinity are obtained.

In another aspect of the invention, it is envisioned that prior to or during recombination or reassortment, polynucleotides generated by the method of the present invention can be subjected to agents or processes which promote the introduction of mutations into the original polynucleotides. The introduction of such mutations would increase the diversity of resulting hybrid polynucleotides and polypeptides encoded therefrom. The agents or processes which promote mutagenesis can include, but are not limited to: (+)-CC-1065, or a synthetic analog such as (+)-CC-1065-(N3-Adenine, see Sun and Hurley, 1992); an N-acetylated or deacetylated 4'-fluoro-4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see, for example, van de Poll et al, 1992); or a N-acetylated or deacetylated 4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see also, van de Poll et al, 1992, pp. 751-758); trivalent chromium, a trivalent chromium salt, a polycyclic aromatic hydrocarbon ("PAH") DNA adduct capable of inhibiting DNA

replication, such as 7-bromomethyl-benz[a]anthracene ("BMA"), tris(2,3-dibromopropyl)phosphate ("Tris-BP"), 1,2-dibromo-3-chloropropane ("DBCP"), 2-bromoacrolein (2BA), benzo[a]pyrene-7,8-dihydrodiol-9-10-epoxide ("BPDE"), a platinum(II) halogen salt, N-hydroxy-2-amino-3-methylimidazo[4,5-f]-quinoline ("N-hydroxy-IQ"), and N-hydroxy-2-amino-1-methyl-6-phenylimidazo[4,5-f]-pyridine ("N-hydroxy-PhIP"). Especially preferred "means for slowing or halting PCR amplification consist of UV light (+)-CC-1065 and (+)-CC-1065-(N3-Adenine). Particularly encompassed means are DNA adducts or polynucleotides comprising the DNA adducts from the polynucleotides or polynucleotides pool, which can be released or removed by a process including heating the solution comprising the polynucleotides prior to further processing.

In another aspect the present invention is directed to a method of producing recombinant proteins having biological activity by treating a sample comprising double-stranded template polynucleotides encoding a wild-type protein under conditions according to the present invention which provide for the production of hybrid or reassorted polynucleotides.

The invention also provides the use of polynucleotide shuffling to shuffle a population of viral genes (e.g., capsid proteins, spike glycoproteins, polymerases, and proteases) or viral genomes (e.g., paramyxoviridae, orthomyxoviridae, herpesviruses, retroviruses, reoviruses and rhinoviruses). In an embodiment, the invention provides a method for shuffling sequences encoding all or portions of immunogenic viral proteins to generate novel combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes which are likely to arise in the natural environment as a consequence of viral evolution; (e.g., such as recombination of influenza virus strains).

The invention also provides a method suitable for shuffling polynucleotide sequences for generating gene therapy vectors and replication-defective gene therapy constructs, such as may be used for human gene therapy, including but not limited to

vaccination vectors for DNA-based vaccination, as well as anti-neoplastic gene therapy and other general therapy formats.

In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

3.1. SATURATION MUTAGENESIS

In one aspect, this invention provides for the use of proprietary codon primers (containing a degenerate N,N,G/T sequence) to introduce point mutations into a polynucleotide, so as to generate a set of progeny polypeptides in which a full range of single amino acid substitutions is represented at each amino acid position. The oligos used are comprised contiguously of a first homologous sequence, a degenerate N,N,G/T sequence, and preferably but not necessarily a second homologous sequence. The downstream progeny translational products from the use of such oligos include all possible amino acid changes at each amino acid site along the polypeptide, because the degeneracy of the N,N,G/T sequence includes codons for all 20 amino acids.

In one aspect, one such degenerate oligo (comprised of one degenerate N,N,G/T cassette) is used for subjecting each original codon in a parental polynucleotide template to a full range of codon substitutions. In another aspect, at least two degenerate N,N,G/T cassettes are used – either in the same oligo or not, for subjecting at least two original codons in a parental polynucleotide template to a full range of codon substitutions. Thus, more than one N,N,G/T sequence can be contained in one oligo to introduce amino acid mutations at more than one site. This plurality of N,N,G/T sequences can be directly contiguous, or separated by one or more additional nucleotide sequence(s). In another aspect, oligos serviceable for introducing additions and deletions can be used either alone or in combination with the codons containing an N,N,G/T sequence, to introduce any combination or permutation of amino acid additions, deletions, and/or substitutions.

In a particular exemplification, it is possible to simultaneously mutagenize two or more contiguous amino acid positions using an oligo that contains contiguous N,N,G/T triplets, i.e. a degenerate (N,N,G/T)_n sequence.

In another aspect, the present invention provides for the use of degenerate cassettes having less degeneracy than the N,N,G/T sequence. For example, it may be desirable in some instances to use (e.g. in an oligo) a degenerate triplet sequence comprised of only one N, where said N can be in the first second or third position of the triplet. Any other

bases including any combinations and permutations thereof can be used in the remaining two positions of the triplet. Alternatively, it may be desirable in some instances to use (e.g. in an oligo) a degenerate N,N,N triplet sequence, or an N,N, G/C triplet sequence.

It is appreciated, however, that the use of a degenerate triplet (such as N,N,G/T or an N,N, G/C triplet sequence) as disclosed in the instant invention is advantageous for several reasons. In one aspect, this invention provides a means to systematically and fairly easily generate the substitution of the full range of possible amino acids (for a total of 20 amino acids) into each and every amino acid position in a polypeptide. Thus, for a 100 amino acid polypeptide, the instant invention provides a way to systematically and fairly easily generate 2000 distinct species (i.e. 20 possible amino acids per position X 100 amino acid positions). It is appreciated that there is provided, through the use of an oligo containing a degenerate N,N,G/T or an N,N, G/C triplet sequence, 32 individual sequences that code for 20 possible amino acids. Thus, in a reaction vessel in which a parental polynucleotide sequence is subjected to saturation mutagenesis using one such oligo, there are generated 32 distinct progeny polynucleotides encoding 20 distinct polypeptides. In contrast, the use of a non-degenerate oligo in site-directed mutagenesis leads to only one progeny polypeptide product per reaction vessel.

This invention also provides for the use of nondegenerate oligos, which can optionally be used in combination with degenerate primers disclosed. It is appreciated that in some situations, it is advantageous to use nondegenerate oligos to generate specific point mutations in a working polynucleotide. This provides a means to generate specific silent point mutations, point mutations leading to corresponding amino acid changes, and point mutations that cause the generation of stop codons and the corresponding expression of polypeptide fragments.

Thus, in a preferred embodiment of this invention, each saturation mutagenesis reaction vessel contains polynucleotides encoding at least 20 progeny polypeptide molecules such that all 20 amino acids are represented at the one specific amino acid position corresponding to the codon position mutagenized in the parental polynucleotide. The 32-fold degenerate progeny polypeptides generated from each saturation mutagenesis reaction vessel can be subjected to clonal amplification (e.g. cloned into a suitable *E. coli*

host using an expression vector) and subjected to expression screening. When an individual progeny polypeptide is identified by screening to display a favorable change in property (when compared to the parental polypeptide), it can be sequenced to identify the correspondingly favorable amino acid substitution contained therein.

It is appreciated that upon mutagenizing each and every amino acid position in a parental polypeptide using saturation mutagenesis as disclosed herein, favorable amino acid changes may be identified at more than one amino acid position. One or more new progeny molecules can be generated that contain a combination of all or part of these favorable amino acid substitutions. For example, if 2 specific favorable amino acid changes are identified in each of 3 amino acid positions in a polypeptide, the permutations include 3 possibilities at each position (no change from the original amino acid, and each of two favorable changes) and 3 positions. Thus, there are $3 \times 3 \times 3$ or 27 total possibilities, including 7 that were previously examined - 6 single point mutations (i.e. 2 at each of three positions) and no change at any position.

In yet another aspect, site-saturation mutagenesis can be used together with shuffling, chimerization, recombination and other mutagenizing processes, along with screening. This invention provides for the use of any mutagenizing process(es), including saturation mutagenesis, in an iterative manner. In one exemplification, the iterative use of any mutagenizing process(es) is used in combination with screening.

Thus, in a non-limiting exemplification, this invention provides for the use of saturation mutagenesis in combination with additional mutagenization processes, such as process where two or more related polynucleotides are introduced into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment.

In addition to performing mutagenesis along the entire sequence of a gene, the instant invention provides that mutagenesis can be use to replace each of any number of bases in a polynucleotide sequence, wherein the number of bases to be mutagenized is preferably every integer from 15 to 100,000. Thus, instead of mutagenizing every position along a molecule, one can subject every a discrete number of bases (preferably a subset

totaling from 15 to 100,000) to mutagenesis. Preferably, a separate nucleotide is used for mutagenizing each position or group of positions along a polynucleotide sequence. A group of 3 positions to be mutagenized may be a codon. The mutations are preferably introduced using a mutagenic primer, containing a heterologous cassette, also referred to as a mutagenic cassette. Preferred cassettes can have from 1 to 500 bases. Each nucleotide position in such heterologous cassettes be N, A, C, G, T, A/C, A/G, A/T, C/G, C/T, G/T, C/G/T, A/G/T, A/C/T, A/C/G, or E, where E is any base that is not A, C, G, or T (E can be referred to as a designer oligo). The tables below show exemplary trinucleotide cassettes (there are over 3000 possibilities in addition to N,N,G/T and N,N,N and N,N,A/C).

In a general sense, saturation mutagenesis is comprised of mutagenizing a complete set of mutagenic cassettes (wherein each cassette is preferably 1-500 bases in length) in defined polynucleotide sequence to be mutagenized (wherein the sequence to be mutagenized is preferably from 15 to 100,000 bases in length). Thusly, a group of mutations (ranging from 1 to 100 mutations) is introduced into each cassette to be mutagenized. A grouping of mutations to be introduced into one cassette can be different or the same from a second grouping of mutations to be introduced into a second cassette during the application of one round of saturation mutagenesis. Such groupings are exemplified by deletions, additions, groupings of particular codons, and groupings of particular nucleotide cassettes.

Defined sequences to be mutagenized (see Fig. 20) include preferably a whole gene, pathway, cDNA, an entire open reading frame (ORF), and entire promoter, enhancer, repressor/transactivator, origin of replication, intron, operator, or any polynucleotide functional group. Generally, a preferred "defined sequences" for this purpose may be any polynucleotide that a 15 base-polynucleotide sequence, and polynucleotide sequences of lengths between 15 bases and 15,000 bases (this invention specifically names every integer in between). Considerations in choosing groupings of codons include types of amino acids encoded by a degenerate mutagenic cassette.

In a particularly preferred exemplification a grouping of mutations that can be introduced into a mutagenic cassette (see Tables 1-85), this invention specifically provides

for degenerate codon substitutions (using degenerate oligos) that code for 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20 amino acids at each position, and a library of polypeptides encoded thereby.

3.2. CHIMERIZATIONS

3.2.1 "SHUFFLING"

Nucleic acid shuffling is a method for *in vitro* or *in vivo* homologous recombination of pools of shorter or smaller polynucleotides to produce a polynucleotide or polynucleotides. Mixtures of related nucleic acid sequences or polynucleotides are subjected to sexual PCR to provide random polynucleotides, and reassembled to yield a library or mixed population of recombinant hybrid nucleic acid molecules or polynucleotides.

In contrast to cassette mutagenesis, only shuffling and error-prone PCR allow one to mutate a pool of sequences blindly (without sequence information other than primers).

The advantage of the mutagenic shuffling of this invention over error-prone PCR alone for repeated selection can best be explained with an example from antibody engineering. Consider DNA shuffling as compared with error-prone PCR (not sexual PCR). The initial library of selected pooled sequences can consist of related sequences of diverse origin (i.e. antibodies from naive mRNA) or can be derived by any type of mutagenesis (including shuffling) of a single antibody gene. A collection of selected complementarity determining regions ("CDRs") is obtained after the first round of affinity selection. In the diagram the thick CDRs confer onto the antibody molecule increased affinity for the antigen. Shuffling allows the free combinatorial association of all of the CDR1s with all of the CDR2s with all of the CDR3s, for example.

This method differs from error-prone PCR, in that it is an inverse chain reaction. In error-prone PCR, the number of polymerase start sites and the number of molecules grows exponentially. However, the sequence of the polymerase start sites and the sequence of the molecules remains essentially the same. In contrast, in nucleic acid

reassembly or shuffling of random polynucleotides the number of start sites and the number (but not size) of the random polynucleotides decreases over time. For polynucleotides derived from whole plasmids the theoretical endpoint is a single, large concatemeric molecule.

Since cross-overs occur at regions of homology, recombination will primarily occur between members of the same sequence family. This discourages combinations of CDRs that are grossly incompatible (e.g., directed against different epitopes of the same antigen). It is contemplated that multiple families of sequences can be shuffled in the same reaction. Further, shuffling generally conserves the relative order, such that, for example, CDR1 will not be found in the position of CDR2.

Rare shufflants will contain a large number of the best (eg. highest affinity) CDRs and these rare shufflants may be selected based on their superior affinity.

CDRs from a pool of 100 different selected antibody sequences can be permuted in up to 1006 different ways. This large number of permutations cannot be represented in a single library of DNA sequences. Accordingly, it is contemplated that multiple cycles of DNA shuffling and selection may be required depending on the length of the sequence and the sequence diversity desired.

Error-prone PCR, in contrast, keeps all the selected CDRs in the same relative sequence, generating a much smaller mutant cloud.

The template polynucleotide which may be used in the methods of this invention may be DNA or RNA. It may be of various lengths depending on the size of the gene or shorter or smaller polynucleotide to be recombined or reassembled. Preferably, the template polynucleotide is from 50 bp to 50 kb. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest can be used in the methods of this invention, and in fact have been successfully used.

The template polynucleotide may be obtained by amplification using the PCR reaction (USPN 4,683,202 and USPN 4,683,195) or other amplification or cloning

methods. However, the removal of free primers from the PCR products before subjecting them to pooling of the PCR products and sexual PCR may provide more efficient results. Failure to adequately remove the primers from the original pool before sexual PCR can lead to a low frequency of crossover clones.

The template polynucleotide often should be double-stranded. A double-stranded nucleic acid molecule is recommended to ensure that regions of the resulting single-stranded polynucleotides are complementary to each other and thus can hybridize to form a double-stranded molecule.

It is contemplated that single-stranded or double-stranded nucleic acid polynucleotides having regions of identity to the template polynucleotide and regions of heterology to the template polynucleotide may be added to the template polynucleotide, at this step. It is also contemplated that two different but related polynucleotide templates can be mixed at this step.

The double-stranded polynucleotide template and any added double-or single-stranded polynucleotides are subjected to sexual PCR which includes slowing or halting to provide a mixture of from about 5 bp to 5 kb or more. Preferably the size of the random polynucleotides is from about 10 bp to 1000 bp, more preferably the size of the polynucleotides is from about 20 bp to 500 bp.

Alternatively, it is also contemplated that double-stranded nucleic acid having multiple nicks may be used in the methods of this invention. A nick is a break in one strand of the double-stranded nucleic acid. The distance between such nicks is preferably 5 bp to 5 kb, more preferably between 10 bp to 1000 bp. This can provide areas of self-priming to produce shorter or smaller polynucleotides to be included with the polynucleotides resulting from random primers, for example.

The concentration of any one specific polynucleotide will not be greater than 1% by weight of the total polynucleotides, more preferably the concentration of any one specific nucleic acid sequence will not be greater than 0.1% by weight of the total nucleic acid.

The number of different specific polynucleotides in the mixture will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

At this step single-stranded or double-stranded polynucleotides, either synthetic or natural, may be added to the random double-stranded shorter or smaller polynucleotides in order to increase the heterogeneity of the mixture of polynucleotides.

It is also contemplated that populations of double-stranded randomly broken polynucleotides may be mixed or combined at this step with the polynucleotides from the sexual PCR process and optionally subjected to one or more additional sexual PCR cycles.

Where insertion of mutations into the template polynucleotide is desired, single-stranded or double-stranded polynucleotides having a region of identity to the template polynucleotide and a region of heterology to the template polynucleotide may be added in a 20 fold excess by weight as compared to the total nucleic acid, more preferably the single-stranded polynucleotides may be added in a 10 fold excess by weight as compared to the total nucleic acid.

Where a mixture of different but related template polynucleotides is desired, populations of polynucleotides from each of the templates may be combined at a ratio of less than about 1:100, more preferably the ratio is less than about 1:40. For example, a backcross of the wild-type polynucleotide with a population of mutated polynucleotide may be desired to eliminate neutral mutations (e.g., mutations yielding an insubstantial alteration in the phenotypic property being selected for). In such an example, the ratio of randomly provided wild-type polynucleotides which may be added to the randomly provided sexual PCR cycle hybrid polynucleotides is approximately 1:1 to about 100:1, and more preferably from 1:1 to 40:1.

The mixed population of random polynucleotides are denatured to form single-stranded polynucleotides and then re-annealed. Only those single-stranded polynucleotides having regions of homology with other single-stranded polynucleotides will re-anneal.

The random polynucleotides may be denatured by heating. One skilled in the art could determine the conditions necessary to completely denature the double-stranded nucleic acid. Preferably the temperature is from 80°C to 100°C, more preferably the temperature is from 90°C to 96°C. Other methods which may be used to denature the polynucleotides include pressure (36) and pH.

The polynucleotides may be re-annealed by cooling. Preferably the temperature is from 20°C to 75°C, more preferably the temperature is from 40°C to 65°C. If a high frequency of crossovers is needed based on an average of only 4 consecutive bases of homology, recombination can be forced by using a low annealing temperature, although the process becomes more difficult. The degree of renaturation which occurs will depend on the degree of homology between the population of single-stranded polynucleotides.

Renaturation can be accelerated by the addition of polyethylene glycol ("PEG") or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%.

The annealed polynucleotides are next incubated in the presence of a nucleic acid polymerase and dNTP's (i.e. dATP, dCTP, dGTP and dTTP). The nucleic acid polymerase may be the Klenow fragment, the Taq polymerase or any other DNA polymerase known in the art.

The approach to be used for the assembly depends on the minimum degree of homology that should still yield crossovers. If the areas of identity are large, Taq polymerase can be used with an annealing temperature of between 45-65°C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30°C. One skilled in the art could vary the temperature of annealing to increase the number of cross-overs achieved.

The polymerase may be added to the random polynucleotides prior to annealing, simultaneously with annealing or after annealing.

The cycle of denaturation, renaturation and incubation in the presence of polymerase is referred to herein as shuffling or reassembly of the nucleic acid. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 50 times, more preferably the sequence is repeated from 10 to 40 times.

The resulting nucleic acid is a larger double-stranded polynucleotide of from about 50 bp to about 100 kb, preferably the larger polynucleotide is from 500 bp to 50 kb.

This larger polynucleotide may contain a number of copies of a polynucleotide having the same size as the template polynucleotide in tandem. This concatemeric polynucleotide is then denatured into single copies of the template polynucleotide. The result will be a population of polynucleotides of approximately the same size as the template polynucleotide. The population will be a mixed population where single or double-stranded polynucleotides having an area of identity and an area of heterology have been added to the template polynucleotide prior to shuffling. These polynucleotides are then cloned into the appropriate vector and the ligation mixture used to transform bacteria.

It is contemplated that the single polynucleotides may be obtained from the larger concatemeric polynucleotide by amplification of the single polynucleotide prior to cloning by a variety of methods including PCR (USPN 4,683,195 and USPN 4,683,202), rather than by digestion of the concatemer.

The vector used for cloning is not critical provided that it will accept a polynucleotide of the desired size. If expression of the particular polynucleotide is desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the polynucleotide to allow expression of the polynucleotide in the host cell. Preferred vectors include the pUC series and the pBR series of plasmids.

The resulting bacterial population will include a number of recombinant polynucleotides having random mutations. This mixed population may be tested to identify the desired recombinant polynucleotides. The method of selection will depend on the polynucleotide desired.

For example, if a polynucleotide which encodes a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the portions of the polynucleotides in the population or library may be tested for their ability to bind to the ligand by methods known in the art (i.e. panning, affinity chromatography). If a polynucleotide which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the polynucleotides in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify polynucleotides which confer the desired properties onto the protein.

It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee WI). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

It is further contemplated that a number of cycles of nucleic acid shuffling may be conducted with polynucleotides from a sub-population of the first population, which sub-population contains DNA encoding the desired recombinant protein. In this manner, proteins with even higher binding affinities or enzymatic activity could be achieved.

It is also contemplated that a number of cycles of nucleic acid shuffling may be conducted with a mixture of wild-type polynucleotides and a sub-population of nucleic acid from the first or subsequent rounds of nucleic acid shuffling in order to remove any silent mutations from the sub-population.

Any source of nucleic acid, in purified form can be utilized as the starting nucleic acid. Thus the process may employ DNA or RNA including messenger RNA, which DNA or RNA may be single or double stranded. In addition, a DNA-RNA hybrid which contains one strand of each may be utilized. The nucleic acid sequence may be of various lengths depending on the size of the nucleic acid sequence to be mutated. Preferably the specific nucleic acid sequence is from 50 to 50000 base pairs. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest may be used in the methods of this invention.

The nucleic acid may be obtained from any source, for example, from plasmids such as pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR, see USPN 4,683,202 and USPN 4,683,195). Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Any specific nucleic acid sequence can be used to produce the population of hybrids by the present process. It is only necessary that a small population of hybrid sequences of the specific nucleic acid sequence exist or be created prior to the present process.

The initial small population of the specific nucleic acid sequences having mutations may be created by a number of different methods. Mutations may be created by error-prone PCR. Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. Alternatively, mutations can be introduced into the template polynucleotide by oligonucleotide-directed mutagenesis. In oligonucleotide-directed mutagenesis, a short sequence of the polynucleotide is removed from the polynucleotide using restriction enzyme digestion and is replaced with a synthetic polynucleotide in which various bases have been altered from the original sequence. The polynucleotide sequence can also be altered by chemical mutagenesis. Chemical mutagens include, for example, sodium bisulfite, nitrous acid,

hydroxylamine, hydrazine or formic acid. Other agents which are analogues of nucleotide precursors include nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. Generally, these agents are added to the PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used. Random mutagenesis of the polynucleotide sequence can also be achieved by irradiation with X-rays or ultraviolet light. Generally, plasmid polynucleotides so mutagenized are introduced into *E. coli* and propagated as a pool or library of hybrid plasmids.

Alternatively the small mixed population of specific nucleic acids may be found in nature in that they may consist of different alleles of the same gene or the same gene from different related species (i.e., cognate genes). Alternatively, they may be related DNA sequences found within one species, for example, the immunoglobulin genes.

Once the mixed population of the specific nucleic acid sequences is generated, the polynucleotides can be used directly or inserted into an appropriate cloning vector, using techniques well-known in the art.

The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates of this invention may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids and phagemids are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large polynucleotides.

If the mixed population of the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of hybrids does not increase. Utility can be readily determined by screening expressed polypeptides.

The DNA shuffling method of this invention can be performed blindly on a pool of unknown sequences. By adding to the reassembly mixture oligonucleotides (with ends that are homologous to the sequences being reassembled) any sequence mixture can be incorporated at any specific position into another sequence mixture. Thus, it is contemplated that mixtures of synthetic oligonucleotides, PCR polynucleotides or even whole genes can be mixed into another sequence library at defined positions. The insertion of one sequence (mixture) is independent from the insertion of a sequence in another part of the template. Thus, the degree of recombination, the homology required, and the diversity of the library can be independently and simultaneously varied along the length of the reassembled DNA.

This approach of mixing two genes may be useful for the humanization of antibodies from murine hybridomas. The approach of mixing two genes or inserting alternative sequences into genes may be useful for any therapeutically used protein, for example, interleukin I, antibodies, tPA and growth hormone. The approach may also be useful in any nucleic acid for example, promoters or introns or 3' untranslated region or 5' untranslated regions of genes to increase expression or alter specificity of expression of proteins. The approach may also be used to mutate ribozymes or aptamers.

Shuffling requires the presence of homologous regions separating regions of diversity. Scaffold-like protein structures may be particularly suitable for shuffling. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta-barrel, and the four-helix bundle which are well-known in the art. This shuffling can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

In vitro Shuffling

The equivalents of some standard genetic matings may also be performed by shuffling *in vitro*. For example, a "molecular backcross" can be performed by repeatedly mixing the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example,

for the removal of neutral mutations that affect unselected characteristics (i.e. immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not, an advantage which cannot be achieved by error-prone mutagenesis or cassette mutagenesis methods.

Large, functional genes can be assembled correctly from a mixture of small random polynucleotides. This reaction may be of use for the reassembly of genes from the highly fragmented DNA of fossils. In addition random nucleic acid fragments from fossils may be combined with polynucleotides from similar genes from related species.

It is also contemplated that the method of this invention can be used for the *in vitro* amplification of a whole genome from a single cell as is needed for a variety of research and diagnostic applications. DNA amplification by PCR is in practice limited to a length of about 40 kb. Amplification of a whole genome such as that of *E. coli* (5,000 kb) by PCR would require about 250 primers yielding 125 forty kb polynucleotides. This approach is not practical due to the unavailability of sufficient sequence data. On the other hand, random production of polynucleotides of the genome with sexual PCR cycles, followed by gel purification of small polynucleotides will provide a multitude of possible primers. Use of this mix of random small polynucleotides as primers in a PCR reaction alone or with the whole genome as the template should result in an inverse chain reaction with the theoretical endpoint of a single concatamer containing many copies of the genome.

100 fold amplification in the copy number and an average polynucleotide size of greater than 50 kb may be obtained when only random polynucleotides are used. It is thought that the larger concatamer is generated by overlap of many smaller polynucleotides. The quality of specific PCR products obtained using synthetic primers will be indistinguishable from the product obtained from unamplified DNA. It is expected that this approach will be useful for the mapping of genomes.

The polynucleotide to be shuffled can be produced as random or non-random polynucleotides, at the discretion of the practitioner. Moreover, this invention provides a method of shuffling that is applicable to a wide range of polynucleotide sizes and types,

including the step of generating polynucleotide monomers to be used as building blocks in the reassembly of a larger polynucleotide. For example, the building blocks can be fragments of genes or they can be comprised of entire genes or gene pathways, or any combination thereof.

In vivo Shuffling

In an embodiment of *in vivo* shuffling, the mixed population of the specific nucleic acid sequence is introduced into bacterial or eukaryotic cells under conditions such that at least two different nucleic acid sequences are present in each host cell. The polynucleotides can be introduced into the host cells by a variety of different methods. The host cells can be transformed with the smaller polynucleotides using methods known in the art, for example treatment with calcium chloride. If the polynucleotides are inserted into a phage genome, the host cell can be transfected with the recombinant phage genome having the specific nucleic acid sequences. Alternatively, the nucleic acid sequences can be introduced into the host cell using electroporation, transfection, lipofection, biolistics, conjugation, and the like.

In general, in this embodiment, the specific nucleic acids sequences will be present in vectors which are capable of stably replicating the sequence in the host cell. In addition, it is contemplated that the vectors will encode a marker gene such that host cells having the vector can be selected. This ensures that the mutated specific nucleic acid sequence can be recovered after introduction into the host cell. However, it is contemplated that the entire mixed population of the specific nucleic acid sequences need not be present on a vector sequence. Rather only a sufficient number of sequences need be cloned into vectors to ensure that after introduction of the polynucleotides into the host cells each host cell contains one vector having at least one specific nucleic acid sequence present therein. It is also contemplated that rather than having a subset of the population of the specific nucleic acids sequences cloned into vectors, this subset may be already stably integrated into the host cell.

It has been found that when two polynucleotides which have regions of identity are inserted into the host cells homologous recombination occurs between the two

polynucleotides. Such recombination between the two mutated specific nucleic acid sequences will result in the production of double or triple hybrids in some situations.

It has also been found that the frequency of recombination is increased if some of the mutated specific nucleic acid sequences are present on linear nucleic acid molecules. Therefore, in a preferred embodiment, some of the specific nucleic acid sequences are present on linear polynucleotides.

After transformation, the host cell transformants are placed under selection to identify those host cell transformants which contain mutated specific nucleic acid sequences having the qualities desired. For example, if increased resistance to a particular drug is desired then the transformed host cells may be subjected to increased concentrations of the particular drug and those transformants producing mutated proteins able to confer increased drug resistance will be selected. If the enhanced ability of a particular protein to bind to a receptor is desired, then expression of the protein can be induced from the transformants and the resulting protein assayed in a ligand binding assay by methods known in the art to identify that subset of the mutated population which shows enhanced binding to the ligand. Alternatively, the protein can be expressed in another system to ensure proper processing.

Once a subset of the first recombined specific nucleic acid sequences (daughter sequences) having the desired characteristics are identified, they are then subject to a second round of recombination.

In the second cycle of recombination, the recombined specific nucleic acid sequences may be mixed with the original mutated specific nucleic acid sequences (parent sequences) and the cycle repeated as described above. In this way a set of second recombined specific nucleic acids sequences can be identified which have enhanced characteristics or encode for proteins having enhanced properties. This cycle can be repeated a number of times as desired.

It is also contemplated that in the second or subsequent recombination cycle, a backcross can be performed. A molecular backcross can be performed by mixing the

desired specific nucleic acid sequences with a large number of the wild-type sequence, such that at least one wild-type nucleic acid sequence and a mutated nucleic acid sequence are present in the same host cell after transformation. Recombination with the wild-type specific nucleic acid sequence will eliminate those neutral mutations that may affect unselected characteristics such as immunogenicity but not the selected characteristics.

In another embodiment of this invention, it is contemplated that during the first round a subset of the specific nucleic acid sequences can be generated as smaller polynucleotides by slowing or halting their PCR amplification prior to introduction into the host cell. The size of the polynucleotides must be large enough to contain some regions of identity with the other sequences so as to homologously recombine with the other sequences. The size of the polynucleotides will range from 0.03 kb to 100 kb more preferably from 0.2 kb to 10 kb. It is also contemplated that in subsequent rounds, all of the specific nucleic acid sequences other than the sequences selected from the previous round may be utilized to generate PCR polynucleotides prior to introduction into the host cells.

The shorter polynucleotide sequences can be single-stranded or double-stranded. If the sequences were originally single-stranded and have become double-stranded they can be denatured with heat, chemicals or enzymes prior to insertion into the host cell. The reaction conditions suitable for separating the strands of nucleic acid are well known in the art.

The steps of this process can be repeated indefinitely, being limited only by the number of possible hybrids which can be achieved. After a certain number of cycles, all possible hybrids will have been achieved and further cycles are redundant.

In an embodiment the same mutated template nucleic acid is repeatedly recombined and the resulting recombinants selected for the desired characteristic.

Therefore, the initial pool or population of mutated template nucleic acid is cloned into a vector capable of replicating in a bacteria such as *E. coli*. The particular vector is not essential, so long as it is capable of autonomous replication in *E. coli*. In a preferred embodiment, the vector is designed to allow the expression and production of any protein

encoded by the mutated specific nucleic acid linked to the vector. It is also preferred that the vector contain a gene encoding for a selectable marker.

The population of vectors containing the pool of mutated nucleic acid sequences is introduced into the *E. coli* host cells. The vector nucleic acid sequences may be introduced by transformation, transfection or infection in the case of phage. The concentration of vectors used to transform the bacteria is such that a number of vectors is introduced into each cell. Once present in the cell, the efficiency of homologous recombination is such that homologous recombination occurs between the various vectors. This results in the generation of hybrids (daughters) having a combination of mutations which differ from the original parent mutated sequences.

The host cells are then clonally replicated and selected for the marker gene present on the vector. Only those cells having a plasmid will grow under the selection.

The host cells which contain a vector are then tested for the presence of favorable mutations. Such testing may consist of placing the cells under selective pressure, for example, if the gene to be selected is an improved drug resistance gene. If the vector allows expression of the protein encoded by the mutated nucleic acid sequence, then such selection may include allowing expression of the protein so encoded, isolation of the protein and testing of the protein to determine whether, for example, it binds with increased efficiency to the ligand of interest.

Once a particular daughter mutated nucleic acid sequence has been identified which confers the desired characteristics, the nucleic acid is isolated either already linked to the vector or separated from the vector. This nucleic acid is then mixed with the first or parent population of nucleic acids and the cycle is repeated.

It has been shown that by this method nucleic acid sequences having enhanced desired properties can be selected.

In an alternate embodiment, the first generation of hybrids are retained in the cells and the parental mutated sequences are added again to the cells. Accordingly, the first cycle of Embodiment I is conducted as described above. However, after the daughter

nucleic acid sequences are identified, the host cells containing these sequences are retained.

The parent mutated specific nucleic acid population, either as polynucleotides or cloned into the same vector is introduced into the host cells already containing the daughter nucleic acids. Recombination is allowed to occur in the cells and the next generation of recombinants, or granddaughters are selected by the methods described above.

This cycle can be repeated a number of times until the nucleic acid or peptide having the desired characteristics is obtained. It is contemplated that in subsequent cycles, the population of mutated sequences which are added to the preferred hybrids may come from the parental hybrids or any subsequent generation.

In an alternative embodiment, the invention provides a method of conducting a "molecular" backcross of the obtained recombinant specific nucleic acid in order to eliminate any neutral mutations. Neutral mutations are those mutations which do not confer onto the nucleic acid or peptide the desired properties. Such mutations may however confer on the nucleic acid or peptide undesirable characteristics. Accordingly, it is desirable to eliminate such neutral mutations. The method of this invention provide a means of doing so.

In this embodiment, after the hybrid nucleic acid, having the desired characteristics, is obtained by the methods of the embodiments, the nucleic acid, the vector having the nucleic acid or the host cell containing the vector and nucleic acid is isolated.

The nucleic acid or vector is then introduced into the host cell with a large excess of the wild-type nucleic acid. The nucleic acid of the hybrid and the nucleic acid of the wild-type sequence are allowed to recombine. The resulting recombinants are placed under the same selection as the hybrid nucleic acid. Only those recombinants which retained the desired characteristics will be selected. Any silent mutations which do not provide the desired characteristics will be lost through recombination with the wild-type

DNA. This cycle can be repeated a number of times until all of the silent mutations are eliminated.

Thus the methods of this invention can be used in a molecular backcross to eliminate unnecessary or silent mutations.

3.2.2. EXONUCLEASE-MEDIATED REASSEMBLY

In a particular embodiment, this invention provides for a method for shuffling, assembling, reassembling, recombining, &/or concatenating at least two polynucleotides to form a progeny polynucleotide (e.g. a chimeric progeny polynucleotide that can be expressed to produce a polypeptide or a gene pathway). In a particular embodiment, a double stranded polynucleotide end (e.g. two single stranded sequences hybridized to each other as hybridization partners) is treated with an exonuclease to liberate nucleotides from one of the two strands, leaving the remaining strand free of its original partner so that, if desired, the remaining strand may be used to achieve hybridization to another partner.

In a particular aspect, a double stranded polynucleotide end (that may be part of - or connected to - a polynucleotide or a nonpolynucleotide sequence) is subjected to a source of exonuclease activity. Serviceable sources of exonuclease activity may be an enzyme with 3' exonuclease activity, an enzyme with 5' exonuclease activity, an enzyme with both 3' exonuclease activity and 5' exonuclease activity, and any combination thereof. An exonuclease can be used to liberate nucleotides from one or both ends of a linear double stranded polynucleotide, and from one to all ends of a branched polynucleotide having more than two ends. The mechanism of action of this liberation is believed to be comprised of an enzymatically-catalyzed hydrolysis of terminal nucleotides, and can be allowed to proceed in a time-dependent fashion, allowing experimental control of the progression of the enzymatic process.

By contrast, a non-enzymatic step may be used to shuffle, assemble, reassemble, recombine, and/or concatenate polynucleotide building blocks that is comprised of subjecting a working sample to denaturing (or "melting") conditions (for example, by changing temperature, pH, and /or salinity conditions) so as to melt a working set of double stranded polynucleotides into single polynucleotide strands. For shuffling, it is

desirable that the single polynucleotide strands participate to some extent in annealment with different hybridization partners (i.e. and not merely revert to exclusive reannealment between what were former partners before the denaturation step). The presence of the former hybridization partners in the reaction vessel, however, does not preclude, and may sometimes even favor, reannealment of a single stranded polynucleotide with its former partner, to recreate an original double stranded polynucleotide.

In contrast to this non-enzymatic shuffling step comprised of subjecting double stranded polynucleotide building blocks to denaturation, followed by annealment, the instant invention further provides an exonuclease-based approach requiring no denaturation – rather, the avoidance of denaturing conditions and the maintenance of double stranded polynucleotide substrates in annealed (i.e. non-denatured) state are necessary conditions for the action of exonucleases (e.g., exonuclease III and red alpha gene product). Additionally in contrast, the generation of single stranded polynucleotide sequences capable of hybridizing to other single stranded polynucleotide sequences is the result of covalent cleavage – and hence sequence destruction - in one of the hybridization partners. For example, an exonuclease III enzyme may be used to enzymatically liberate 3' terminal nucleotides in one hybridization strand (to achieve covalent hydrolysis in that polynucleotide strand); and this favors hybridization of the remaining single strand to a new partner (since its former partner was subjected to covalent cleavage).

By way of further illustration, a specific exonuclease, namely exonuclease III is provided herein as an example of a 3' exonuclease; however, other exonucleases may also be used, including enzymes with 5' exonuclease activity and enzymes with 3' exonuclease activity, and including enzymes not yet discovered and enzymes not yet developed. It is particularly appreciated that enzymes can be discovered, optimized (e.g. engineered by directed evolution), or both discovered and optimized specifically for the instantly disclosed approach that have more optimal rates &/or more highly specific activities &/or greater lack of unwanted activities. In fact it is expected that the instant invention may encourage the discovery &/or development of such designer enzymes. In sum, this invention may be practiced with a variety of currently available exonuclease enzymes, as well as enzymes not yet discovered and enzymes not yet developed.

The exonuclease action of exonuclease III requires a working double stranded polynucleotide end that is either blunt or has a 5' overhang, and the exonuclease action is comprised of enzymatically liberating 3' terminal nucleotides, leaving a single stranded 5' end that becomes longer and longer as the exonuclease action proceeds (see Figure 1). Any 5' overhangs produced by this approach may be used to hybridize to another single stranded polynucleotide sequence (which may also be a single stranded polynucleotide or a terminal overhang of a partially double stranded polynucleotide) that shares enough homology to allow hybridization. The ability of these exonuclease III-generated single stranded sequences (e.g. in 5' overhangs) to hybridize to other single stranded sequences allows two or more polynucleotides to be shuffled, assembled, reassembled, &/or concatenated.

Furthermore, it is appreciated that one can protect the end of a double stranded polynucleotide or render it susceptible to a desired enzymatic action of a serviceable exonuclease as necessary. For example, a double stranded polynucleotide end having a 3' overhang is not susceptible to the exonuclease action of exonuclease III. However, it may be rendered susceptible to the exonuclease action of exonuclease III by a variety of means; for example, it may be blunted by treatment with a polymerase, cleaved to provide a blunt end or a 5' overhang, joined (ligated or hybridized) to another double stranded polynucleotide to provide a blunt end or a 5' overhang, hybridized to a single stranded polynucleotide to provide a blunt end or a 5' overhang, or modified by any of a variety of means).

According to one aspect, an exonuclease may be allowed to act on one or on both ends of a linear double stranded polynucleotide and proceed to completion, to near completion, or to partial completion. When the exonuclease action is allowed to go to completion, the result will be that the length of each 5' overhang will extend far towards the middle region of the polynucleotide in the direction of what might be considered a "rendezvous point" (which may be somewhere near the polynucleotide midpoint). Ultimately, this results in the production of single stranded polynucleotides (that can become dissociated) that are each about half the length of the original double stranded polynucleotide (see Figure 1). Alternatively, an exonuclease-mediated reaction can be terminated before proceeding to completion.

Thus this exonuclease-mediated approach is serviceable for shuffling, assembling &/or reassembling, recombining, and concatenating polynucleotide building blocks, which polynucleotide building blocks can be up to ten bases long or tens of bases long or hundreds of bases long or thousands of bases long or tens of thousands of bases long or hundreds of thousands of bases long or millions of bases long or even longer.

This exonuclease-mediated approach is based on the action of double stranded DNA specific exodeoxyribonuclease activity of *E. coli* exonuclease III. Substrates for exonuclease III may be generated by subjecting a double stranded polynucleotide to fragmentation. Fragmentation may be achieved by mechanical means (e.g., shearing, sonication, etc.), by enzymatic means (e.g. using restriction enzymes), and by any combination thereof. Fragments of a larger polynucleotide may also be generated by polymerase-mediated synthesis.

Exonuclease III is a 28K monomeric enzyme, product of the *xthA* gene of *E. coli* with four known activities: exodeoxyribonuclease (alternatively referred to as exonuclease herein), RNaseH, DNA-3'-phosphatase, and AP endonuclease. The exodeoxyribonuclease activity is specific for double stranded DNA. The mechanism of action is thought to involve enzymatic hydrolysis of DNA from a 3' end progressively towards a 5' direction, with formation of nucleoside 5'-phosphates and a residual single strand. The enzyme does not display efficient hydrolysis of single stranded DNA, single-stranded RNA, or double-stranded RNA; however it degrades RNA in an DNA-RNA hybrid releasing nucleoside 5'-phosphates. The enzyme also releases inorganic phosphate specifically from 3'phosphomonoester groups on DNA, but not from RNA or short oligonucleotides. Removal of these groups converts the terminus into a primer for DNA polymerase action.

Additional examples of enzymes with exonuclease activity include red-alpha and venom phosphodiesterases. Red alpha (*red* gene product (also referred to as lambda exonuclease) is of bacteriophage origin. The *red* gene is transcribed from the leftward promoter and its product is involved (24 kD) in recombination. Red alpha gene product acts processively from 5'-phosphorylated termini to liberate mononucleotides from duplex

DNA (Takahashi & Kobayashi, 1990). Venom phosphodiesterases (Laskowski, 1980) is capable of rapidly opening supercoiled DNA.

3.2.3. NON-STOCHASTIC LIGATION REASSEMBLY

In one aspect, the present invention provides a non-stochastic method termed synthetic ligation reassembly (SLR), that is somewhat related to stochastic shuffling, save that the nucleic acid building blocks are not shuffled or concatenated or chimerized randomly, but rather are assembled non-stochastically.

A particularly glaring difference is that the instant SLR method does not depend on the presence of a high level of homology between polynucleotides to be shuffled. In contrast, prior methods, particularly prior stochastic shuffling methods require that presence of a high level of homology, particularly at coupling sites, between polynucleotides to be shuffled. Accordingly these prior methods favor the regeneration of the original progenitor molecules, and are suboptimal for generating large numbers of novel progeny chimeras, particularly full-length progenies. The instant invention, on the other hand, can be used to non-stochastically generate libraries (or sets) of progeny molecules comprised of over 10^{100} different chimeras. Conceivably, SLR can even be used to generate libraries comprised of over 10^{1000} different progeny chimeras with (no upper limit in sight).

Thus, in one aspect, the present invention provides a method, which method is non-stochastic, of producing a set of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design, which method is comprised of the steps of generating by design a plurality of specific nucleic acid building blocks having serviceable mutually compatible ligatable ends, and assembling these nucleic acid building blocks, such that a designed overall assembly order is achieved.

The mutually compatible ligatable ends of the nucleic acid building blocks to be assembled are considered to be "serviceable" for this type of ordered assembly if they enable the building blocks to be coupled in predetermined orders. Thus, in one aspect, the overall assembly order in which the nucleic acid building blocks can be coupled is

specified by the design of the ligatable ends and, if more than one assembly step is to be used, then the overall assembly order in which the nucleic acid building blocks can be coupled is also specified by the sequential order of the assembly step(s). Figure 4, Panel C illustrates an exemplary assembly process comprised of 2 sequential steps to achieve a designed (non-stochastic) overall assembly order for five nucleic acid building blocks. In a preferred embodiment of this invention, the annealed building pieces are treated with an enzyme, such as a ligase (e.g. T4 DNA ligase), achieve covalent bonding of the building pieces.

In a preferred embodiment, the design of nucleic acid building blocks is obtained upon analysis of the sequences of a set of progenitor nucleic acid templates that serve as a basis for producing a progeny set of finalized chimeric nucleic acid molecules. These progenitor nucleic acid templates thus serve as a source of sequence information that aids in the design of the nucleic acid building blocks that are to be mutagenized, i.e. chimerized or shuffled.

In one exemplification, this invention provides for the chimerization of a family of related genes and their encoded family of related products. In a particular exemplification, the encoded products are enzymes. As a representative list of families of enzymes which may be mutagenized in accordance with the aspects of the present invention, there may be mentioned, the following enzymes and their functions:

1 Lipase/Esterase

- a. Enantioselective hydrolysis of esters (lipids)/ thioesters
 - 1) Resolution of racemic mixtures
 - 2) Synthesis of optically active acids or alcohols from *meso*-diesters
- b. Selective syntheses
 - 1) Regiospecific hydrolysis of carbohydrate esters
 - 2) Selective hydrolysis of cyclic secondary alcohols
- c. Synthesis of optically active esters, lactones, acids, alcohols
 - 1) Transesterification of activated/nonactivated esters
 - 2) Interesterification
 - 3) Optically active lactones from hydroxyesters

- 4) Regio- and enantioselective ring opening of anhydrides
 - d. Detergents
 - e. Fat/Oil conversion
 - f. Cheese ripening
2. **Protease**
- a. Ester/amide synthesis
 - b. Peptide synthesis
 - c. Resolution of racemic mixtures of amino acid esters
 - d. Synthesis of non-natural amino acids
 - e. Detergents/protein hydrolysis
3. **Glycosidase/Glycosyl transferase**
- a. Sugar/polymer synthesis
 - b. Cleavage of glycosidic linkages to form mono, di- and oligosaccharides
 - c. Synthesis of complex oligosaccharides
 - d. Glycoside synthesis using UDP-galactosyl transferase
 - e. Transglycosylation of disaccharides, glycosyl fluorides, aryl galactosides
 - f. Glycosyl transfer in oligosaccharide synthesis
 - g. Diastereoselective cleavage of -glucosylsulfoxides
 - h. Asymmetric glycosylations
 - i. Food processing
 - j. Paper processing
4. **Phosphatase/Kinase**
- a. Synthesis/hydrolysis of phosphate esters
 - 1) Regio-, enantioselective phosphorylation
 - 2) Introduction of phosphate esters
 - 3) Synthesize phospholipid precursors
 - 4) Controlled polynucleotide synthesis
 - b. Activate biological molecule
 - c. Selective phosphate bond formation without protecting groups

- 5 **Mono/Dioxygenase**
 - a. Direct oxyfunctionalization of unactivated organic substrates
 - b. Hydroxylation of alkane, aromatics, steroids
 - c. Epoxidation of alkenes
 - d. Enantioselective sulfoxidation
 - e. Regio- and stereoselective Bayer-Villiger oxidations

- 6 **Haloperoxidase**
 - a. Oxidative addition of halide ion to nucleophilic sites
 - b. Addition of hypohalous acids to olefinic bonds
 - c. Ring cleavage of cyclopropanes
 - d. Activated aromatic substrates converted to *ortho* and *para* derivatives
 - e. 1,3 diketones converted to 2-halo-derivatives
 - f. Heteroatom oxidation of sulfur and nitrogen containing substrates
 - g. Oxidation of enol acetates, alkynes and activated aromatic rings

- 7 **Lignin peroxidase/Diarylpropane peroxidase**
 - a. Oxidative cleavage of C-C bonds
 - b. Oxidation of benzylic alcohols to aldehydes
 - c. Hydroxylation of benzylic carbons
 - d. Phenol dimerization
 - e. Hydroxylation of double bonds to form diols
 - f. Cleavage of lignin aldehydes

- 8 **Epoxide hydrolase**
 - a. Synthesis of enantiomerically pure bioactive compounds
 - b. Regio- and enantioselective hydrolysis of epoxide
 - c. Aromatic and olefinic epoxidation by monooxygenases to form epoxides
 - d. Resolution of racemic epoxides
 - e. Hydrolysis of steroid epoxides

- 9 **Nitrile hydratase/nitrilase**
 - a. Hydrolysis of aliphatic nitriles to carboxamides

- b. Hydrolysis of aromatic, heterocyclic, unsaturated aliphatic nitriles to corresponding acids
 - c. Hydrolysis of acrylonitrile
 - d. Production of aromatic and carboxamides, carboxylic acids (nicotinamide, picolinamide, isonicotinamide)
 - e. Regioselective hydrolysis of acrylic dinitrile
 - f. - amino acids from hydroxynitriles
- 10 Transaminase**
- a. Transfer of amino groups into oxo-acids
- 11 Amidase/Acylase**
- a. Hydrolysis of amides, amidines, and other C-N bonds
 - b. Non-natural amino acid resolution and synthesis

These exemplifications, while illustrating certain specific aspects of the invention, do not portray the limitations or circumscribe the scope of the disclosed invention.

Thus according to one aspect of this invention, the sequences of a plurality of progenitor nucleic acid templates are aligned in order to select one or more demarcation points, which demarcation points can be located at an area of homology, and are comprised of one or more nucleotides, and which demarcation points are shared by at least two of the progenitor templates. The demarcation points can be used to delineate the boundaries of nucleic acid building blocks to be generated. Thus, the demarcation points identified and selected in the progenitor molecules serve as potential chimerization points in the assembly of the progeny molecules.

Preferably a serviceable demarcation point is an area of homology (comprised of at least one homologous nucleotide base) shared by at least two progenitor templates. More preferably a serviceable demarcation point is an area of homology that is shared by at least half of the progenitor templates. More preferably still a serviceable demarcation point is an area of homology that is shared by at least two thirds of the progenitor templates. Even more preferably a serviceable demarcation points is an area of homology that is shared by

at least three fourths of the progenitor templates. Even more preferably still a serviceable demarcation points is an area of homology that is shared by at almost all of the progenitor templates. Even more preferably still a serviceable demarcation point is an area of homology that is shared by all of the progenitor templates.

The process of designing nucleic acid building blocks and of designing the mutually compatible ligatable ends of the nucleic acid building blocks to be assembled is illustrated in Figures 6 and 7. As shown, the alignment of a set of progenitor templates reveals several naturally occurring demarcation points, and the identification of demarcation points shared by these templates helps to non-stochastically determine the building blocks to be generated and used for the generation of the progeny chimeric molecules.

In a preferred embodiment, this invention provides that the ligation reassembly process is performed exhaustively in order to generate an exhaustive library. In other words, all possible ordered combinations of the nucleic acid building blocks are represented in the set of finalized chimeric nucleic acid molecules. At the same time, in a particularly preferred embodiment, the assembly order (i.e. the order of assembly of each building block in the 5' to 3' sequence of each finalized chimeric nucleic acid) in each combination is by design (or non-stochastic). Because of the non-stochastic nature of this invention, the possibility of unwanted side products is greatly reduced.

In another preferred embodiment, this invention provides that the ligation reassembly process is performed systematically, for example in order to generate a systematically compartmentalized library, with compartments that can be screened systematically, e.g. one by one. In other words this invention provides that, through the selective and judicious use of specific nucleic acid building blocks, coupled with the selective and judicious use of sequentially stepped assembly reactions, an experimental design can be achieved where specific sets of progeny products are made in each of several reaction vessels. This allows a systematic examination and screening procedure to be performed. Thus, it allows a potentially very large number of progeny molecules to be examined systematically in smaller groups.

Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, the instant invention provides for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated preferably comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design. In a particularly preferred embodiment of this invention, such a generated library is comprised of preferably greater than 10^3 different progeny molecular species, more preferably greater than 10^5 different progeny molecular species, more preferably still greater than 10^{10} different progeny molecular species, more preferably still greater than 10^{15} different progeny molecular species, more preferably still greater than 10^{20} different progeny molecular species, more preferably still greater than 10^{30} different progeny molecular species, more preferably still greater than 10^{40} different progeny molecular species, more preferably still greater than 10^{50} different progeny molecular species, more preferably still greater than 10^{60} different progeny molecular species, more preferably still greater than 10^{70} different progeny molecular species, more preferably still greater than 10^{80} different progeny molecular species, more preferably still greater than 10^{100} different progeny molecular species, more preferably still greater than 10^{110} different progeny molecular species, more preferably still greater than 10^{120} different progeny molecular species, more preferably still greater than 10^{130} different progeny molecular species, more preferably still greater than 10^{140} different progeny molecular species, more preferably still greater than 10^{150} different progeny molecular species, more preferably still greater than 10^{175} different progeny molecular species, more preferably still greater than 10^{200} different progeny molecular species, more preferably still greater than 10^{300} different progeny molecular species, more preferably still greater than 10^{400} different progeny molecular species, more preferably still greater than 10^{500} different progeny molecular species, and even more preferably still greater than 10^{1000} different progeny molecular species.

In one aspect, a set of finalized chimeric nucleic acid molecules, produced as described is comprised of a polynucleotide encoding a polypeptide. According to one preferred embodiment, this polynucleotide is a gene, which may be a man-made gene. According to another preferred embodiment, this polynucleotide is a gene pathway, which

may be a man-made gene pathway. This invention provides that one or more man-made genes generated by this invention may be incorporated into a man-made gene pathway, such as a pathway operable in a eukaryotic organism (including a plant).

It is appreciated that the power of this invention is exceptional, as there is much freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This invention provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecularly homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another exemplification, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g. one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an in vitro process (e.g. by mutagenesis) or in an in vivo process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

Thus, according to another embodiment, this invention provides that a nucleic acid building block can be used to introduce an intron. Thus, this invention provides that functional introns may be introduced into a man-made gene of this invention. This invention also provides that functional introns may be introduced into a man-made gene pathway of this invention. Accordingly, this invention provides for the generation of a

chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced intron(s).

Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced intron(s). Preferably, the artificially introduced intron(s) are functional in one or more host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing. This invention provides a process of producing man-made intron-containing polynucleotides to be introduced into host organisms for recombination and/or splicing.

The ability to achieve chimerizations, using couplings as described herein, in areas of little or no homology among the progenitor molecules, is particularly useful, and in fact critical, for the assembly of novel gene pathways. This invention thus provides for the generation of novel man-made gene pathways using synthetic ligation reassembly. In a particular aspect, this is achieved by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, to confer operability to a novel gene pathway when it is introduced into the intended host. In a particular exemplification, this invention provides for the generation of novel man-made gene pathways that is operable in a plurality of intended hosts (e.g. in a microbial organism as well as in a plant cell). This can be achieved, for example, by the introduction of a plurality of regulatory sequences, comprised of a regulatory sequence that is operable in a first intended host and a regulatory sequence that is operable in a second intended host. A similar process can be performed to achieve operability of a gene pathway in a third intended host species, etc. The number of intended host species can be each integer from 1 to 10 or alternatively over 10. Alternatively, for example, operability of a gene pathway in a plurality of intended hosts can be achieved by the introduction of a regulatory sequence having intrinsic operability in a plurality of intended hosts.

Thus, according to a particular embodiment, this invention provides that a nucleic acid building block can be used to introduce a regulatory sequence, particularly a regulatory sequence for gene expression. Preferred regulatory sequences include, but are not limited to, those that are man-made, and those found in archeal, bacterial, eukaryotic

(including mitochondrial), viral, and prionic or prion-like organisms. Preferred regulatory sequences include but are not limited to, promoters, operators, and activator binding sites. Thus, this invention provides that functional regulatory sequences may be introduced into a man-made gene of this invention. This invention also provides that functional regulatory sequences may be introduced into a man-made gene pathway of this invention.

Accordingly, this invention provides for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced regulatory sequence(s). Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced regulatory sequence(s). Preferably, an artificially introduced regulatory sequence(s) is operatively linked to one or more genes in the man-made polynucleotide, and are functional in one or more host cells.

Preferred bacterial promoters that are serviceable for this invention include *lacI*, *lacZ*, T3, T7, *gpt*, *lambda P_R*, *P_L* and *trp*. Serviceable eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Particular plant regulatory sequences include promoters active in directing transcription in plants, either constitutively or stage and/or tissue specific, depending on the use of the plant or parts thereof. These promoters include, but are not limited to promoters showing constitutive expression, such as the 35S promoter of Cauliflower Mosaic Virus (CaMV) (Guilley et al., 1982), those for leaf-specific expression, such as the promoter of the ribulose biphosphate carboxylase small subunit gene (Coruzzi et al., 1984), those for root-specific expression, such as the promoter from the glutamin synthase gene (Tingey et al., 1987), those for seed-specific expression, such as the cruciferin A promoter from *Brassica napus* (Ryan et al., 1989), those for tuber-specific expression, such as the class-I patatin promoter from potato (Rocha-Sasa et al., 1989; Wenzler et al., 1989) or those for fruit-specific expression, such as the polygalacturonase (PG) promoter from tomato (Bird et al., 1988).

Other regulatory sequences that are preferred for this invention include terminator sequences and polyadenylation signals and any such sequence functioning as such in plants, the choice of which is within the level of the skilled artisan. An example of such

sequences is the 3' flanking region of the nopaline synthase (nos) gene of *Agrobacterium tumefaciens* (Bevan, 1984). The regulatory sequences may also include enhancer sequences, such as found in the 35S promoter of CaMV, and mRNA stabilizing sequences such as the leader sequence of Alfalfa Mosaic Virus (AIMV) RNA4 (Brederode et al., 1980) or any other sequences functioning in a like manner.

Man-made genes produced using this invention can also serve as a substrate for recombination with another nucleic acid. Likewise, a man-made gene pathway produced using this invention can also serve as a substrate for recombination with another nucleic acid. In a preferred instance, the recombination is facilitated by, or occurs at, areas of homology between the man-made intron-containing gene and a nucleic acid which serves as a recombination partner. In a particularly preferred instance, the recombination partner may also be a nucleic acid generated by this invention, including a man-made gene or a man-made gene pathway. Recombination may be facilitated by or may occur at areas of homology that exist at the one (or more) artificially introduced intron(s) in the man-made gene.

The synthetic ligation reassembly method of this invention utilizes a plurality of nucleic acid building blocks, each of which preferably has two ligatable ends. The two ligatable ends on each nucleic acid building block may be two blunt ends (i.e. each having an overhang of zero nucleotides), or preferably one blunt end and one overhang, or more preferably still two overhangs.

A serviceable overhang for this purpose may be a 3' overhang or a 5' overhang. Thus, a nucleic acid building block may have a 3' overhang or alternatively a 5' overhang or alternatively two 3' overhangs or alternatively two 5' overhangs. The overall order in which the nucleic acid building blocks are assembled to form a finalized chimeric nucleic acid molecule is determined by purposeful experimental design and is not random.

According to one preferred embodiment, a nucleic acid building block is generated by chemical synthesis of two single-stranded nucleic acids (also referred to as single-stranded oligos) and contacting them so as to allow them to anneal to form a double-stranded nucleic acid building block.

A double-stranded nucleic acid building block can be of variable size. The sizes of these building blocks can be small or large depending on the choice of the experimenter. Preferred sizes for building block range from 1 base pair (not including any overhangs) to 100,000 base pairs (not including any overhangs). Other preferred size ranges are also provided, which have lower limits of from 1 bp to 10,000 bp (including every integer value in between), and upper limits of from 2 bp to 100,000 bp (including every integer value in between).

It is appreciated that current methods of polymerase-based amplification can be used to generate double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length with high fidelity. Chemical synthesis (e.g. phosphoramidite-based) can be used to generate nucleic acids of up to hundreds of nucleotides in length with high fidelity; however, these can be assembled, e.g. using overhangs or sticky ends, to form double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length if so desired.

A combination of methods (e.g. phosphoramidite-based chemical synthesis and PCR) can also be used according to this invention. Thus, nucleic acid building block made by different methods can also be used in combination to generate a progeny molecule of this invention.

The use of chemical synthesis to generate nucleic acid building blocks is particularly preferred in this invention & is advantageous for other reasons as well, including procedural safety and ease. No cloning or harvesting or actual handling of any biological samples is required. The design of the nucleic acid building blocks can be accomplished on paper. Accordingly, this invention teaches an advance in procedural safety in recombinant technologies.

Nonetheless, according to one preferred embodiment, a double-stranded nucleic acid building block according to this invention may also be generated by polymerase-based amplification of a polynucleotide template. In a non-limiting exemplification, as illustrated in Figure 2, a first polymerase-based amplification reaction using a first set of

primers, F_2 and R_1 , is used to generate a blunt-ended product (labeled Reaction 1, Product 1), which is essentially identical to Product A. A second polymerase-based amplification reaction using a second set of primers, F_1 and R_2 , is used to generate a blunt-ended product (labeled Reaction 2, Product 2), which is essentially identical to Product B. These two products are mixed and allowed to melt and anneal, generating potentially useful double-stranded nucleic acid building blocks with two overhangs. In the example of Fig. 2, the product with the 3' overhangs (Product C) is selected by nuclease-based degradation of the other 3 products using a 3' acting exonuclease, such as exonuclease III. It is appreciated that a 5' acting exonuclease (e.g. red alpha) may be also be used, for example to select Product D instead. It is also appreciated that other selection means can also be used, including hybridization-based means, and that these means can incorporate a further means, such as a magnetic bead-based means, to facilitate separation of the desired product.

Many other methods exist by which a double-stranded nucleic acid building block can be generated that is serviceable for this invention; and these are known in the art and can be readily performed by the skilled artisan.

According to particularly preferred embodiment, a double-stranded nucleic acid building block that is serviceable for this invention is generated by first generating two single stranded nucleic acids and allowing them to anneal to form a double-stranded nucleic acid building block. The two strands of a double-stranded nucleic acid building block may be complementary at every nucleotide apart from any that form an overhang; thus containing no mismatches, apart from any overhang(s). According to another embodiment, the two strands of a double-stranded nucleic acid building block are complementary at fewer than every nucleotide apart from any that form an overhang. Thus, according to this embodiment, a double-stranded nucleic acid building block can be used to introduce codon degeneracy. Preferably the codon degeneracy is introduced using the site-saturation mutagenesis described herein, using one or more $N,N,G/T$ cassettes or alternatively using one or more N,N,N cassettes.

Contained within an exemplary experimental design for achieving an ordered assembly according to this invention are:

- 1) The design of specific nucleic acid building blocks.
- 2) The design of specific ligatable ends on each nucleic acid building block.
- 3) The design of a particular order of assembly of the nucleic acid building blocks.

An overhang may be a 3' overhang or a 5' overhang. An overhang may also have a terminal phosphate group or alternatively may be devoid of a terminal phosphate group (having, e.g., a hydroxyl group instead). An overhang may be comprised of any number of nucleotides. Preferably an overhang is comprised of 0 nucleotides (as in a blunt end) to 10,000 nucleotides. Thus, a wide range of overhang sizes may be serviceable. Accordingly, the lower limit may be each integer from 1-200 and the upper limit may be each integer from 2-10,000. According to a particular exemplification, an overhang may consist of anywhere from 1 nucleotide to 200 nucleotides (including every integer value in between).

The final chimeric nucleic acid molecule may be generated by sequentially assembling 2 or more building blocks at a time until all the designated building blocks have been assembled. A working sample may optionally be subjected to a process for size selection or purification or other selection or enrichment process between the performance of two assembly steps. Alternatively, the final chimeric nucleic acid molecule may be generated by assembling all the designated building blocks at once in one step.

Utility

The *in vivo* recombination method of this invention can be performed blindly on a pool of unknown hybrids or alleles of a specific polynucleotide or sequence. However, it is not necessary to know the actual DNA or RNA sequence of the specific polynucleotide.

The approach of using recombination within a mixed population of genes can be useful for the generation of any useful proteins, for example, interleukin I, antibodies, tPA and growth hormone. This approach may be used to generate proteins having altered

specificity or activity. The approach may also be useful for the generation of hybrid nucleic acid sequences, for example, promoter regions, introns, exons, enhancer sequences, 3' untranslated regions or 5' untranslated regions of genes. Thus this approach may be used to generate genes having increased rates of expression. This approach may also be useful in the study of repetitive DNA sequences. Finally, this approach may be useful to mutate ribozymes or aptamers.

Scaffold-like regions separating regions of diversity in proteins may be particularly suitable for the methods of this invention. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta barrel, and the four-helix bundle. The methods of this invention can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

The equivalents of some standard genetic matings may also be performed by the methods of this invention. For example, a "molecular" backcross can be performed by repeated mixing of the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (i.e. immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not.

3.2.4. END-SELECTION

This invention provides a method for selecting a subset of polynucleotides from a starting set of polynucleotides, which method is based on the ability to discriminate one or more selectable features (or selection markers) present anywhere in a working polynucleotide, so as to allow one to perform selection for (positive selection) &/or against (negative selection) each selectable polynucleotide. In a preferred aspect, a method is provided termed end-selection, which method is based on the use of a selection marker located in part or entirely in a terminal region of a selectable polynucleotide, and such a selection marker may be termed an "end-selection marker".

End-selection may be based on detection of naturally occurring sequences or on detection of sequences introduced experimentally (including by any mutagenesis procedure mentioned herein and not mentioned herein) or on both, even within the same polynucleotide. An end-selection marker can be a structural selection marker or a functional selection marker or both a structural and a functional selection marker. An end-selection marker may be comprised of a polynucleotide sequence or of a polypeptide sequence or of any chemical structure or of any biological or biochemical tag, including markers that can be selected using methods based on the detection of radioactivity, of enzymatic activity, of fluorescence, of any optical feature, of a magnetic property (e.g. using magnetic beads), of immunoreactivity, and of hybridization.

End-selection may be applied in combination with any method serviceable for performing mutagenesis. Such mutagenesis methods include, but are not limited to, methods described herein (*supra* and *infra*). Such methods include, by way of non-limiting exemplification, any method that may be referred herein or by others in the art by any of the following terms: "saturation mutagenesis", "shuffling", "recombination", "re-assembly", "error-prone PCR", "assembly PCR", "sexual PCR", "crossover PCR", "oligonucleotide primer-directed mutagenesis", "recursive (&/or exponential) ensemble mutagenesis (see Arkin and Youvan, 1992)", "cassette mutagenesis", "in vivo mutagenesis", and "in vitro mutagenesis". Moreover, end-selection may be performed on molecules produced by any mutagenesis &/or amplification method (see, e.g., Arnold, 1993; Caldwell and Joyce, 1992; Stemmer, 1994; following which method it is desirable to select for (including to screen for the presence of) desirable progeny molecules.

In addition, end-selection may be applied to a polynucleotide apart from any mutagenesis method. In a preferred embodiment, end-selection, as provided herein, can be used in order to facilitate a cloning step, such as a step of ligation to another polynucleotide (including ligation to a vector). This invention thus provides for end-selection as a serviceable means to facilitate library construction, selection &/or enrichment for desirable polynucleotides, and cloning in general.

In a particularly preferred embodiment, end-selection can be based on (positive) selection for a polynucleotide; alternatively end-selection can be based on (negative) selection against a polynucleotide; and alternatively still, end-selection can be based on both (positive) selection for, and on (negative) selection against, a polynucleotide. End-selection, along with other methods of selection &/or screening, can be performed in an iterative fashion, with any combination of like or unlike selection &/or screening methods and serviceable mutagenesis methods, all of which can be performed in an iterative fashion and in any order, combination, and permutation.

It is also appreciated that, according to one embodiment of this invention, end-selection may also be used to select a polynucleotide that is at least in part: circular (e.g. a plasmid or any other circular vector or any other polynucleotide that is partly circular), &/or branched, &/or modified or substituted with any chemical group or moiety. In accord with this embodiment, a polynucleotide may be a circular molecule comprised of an intermediate or central region, which region is flanked on a 5' side by a 5' flanking region (which, for the purpose of end-selection, serves in like manner to a 5' terminal region of a non-circular polynucleotide) and on a 3' side by a 3' terminal region (which, for the purpose of end-selection, serves in like manner to a 3' terminal region of a non-circular polynucleotide). As used in this non-limiting exemplification, there may be sequence overlap between any two regions or even among all three regions.

In one non-limiting aspect of this invention, end-selection of a linear polynucleotide is performed using a general approach based on the presence of at least one end-selection marker located at or near a polynucleotide end or terminus (that can be either a 5' end or a 3' end). In one particular non-limiting exemplification, end-selection is based on selection for a specific sequence at or near a terminus such as, but not limited to, a sequence recognized by an enzyme that recognizes a polynucleotide sequence. An enzyme that recognizes and catalyzes a chemical modification of a polynucleotide is referred to herein as a polynucleotide-acting enzyme. In a preferred embodiment, serviceable polynucleotide-acting enzymes are exemplified non-exclusively by enzymes with polynucleotide-cleaving activity, enzymes with polynucleotide-methylating activity, enzymes with polynucleotide-ligating activity, and enzymes with a plurality of

distinguishable enzymatic activities (including non-exclusively, e.g., both polynucleotide-cleaving activity and polynucleotide-ligating activity).

Relevant polynucleotide-acting enzymes thus also include any commercially available or non-commercially available polynucleotide endonucleases and their companion methylases including those catalogued at the website <http://www.neb.com/rebase>, and those mentioned in the following cited reference (Roberts and Macelis, 1996). Preferred polynucleotide endonucleases include – but are not limited to – type II restriction enzymes (including type IIS), and include enzymes that cleave both strands of a double stranded polynucleotide (e.g. *Not* I, which cleaves both strands at 5'...GC/GGCCGC...3') and enzymes that cleave only one strand of a double stranded polynucleotide, i.e. enzymes that have polynucleotide-nicking activity, (e.g. *N. Bst*NB I, which cleaves only one strand at 5'...GAGTCNNNN/N...3'). Relevant polynucleotide-acting enzymes also include type III restriction enzymes.

It is appreciated that relevant polynucleotide-acting enzymes also include any enzymes that may be developed in the future, though currently unavailable, that are serviceable for generating a ligation compatible end, preferably a sticky end, in a polynucleotide.

In one preferred exemplification, a serviceable selection marker is a restriction site in a polynucleotide that allows a corresponding type II (or type IIS) restriction enzyme to cleave an end of the polynucleotide so as to provide a ligatable end (including a blunt end or alternatively a sticky end with at least a one base overhang) that is serviceable for a desirable ligation reaction without cleaving the polynucleotide internally in a manner that destroys a desired internal sequence in the polynucleotide. Thus it is provided that, among relevant restriction sites, those sites that do not occur internally (i.e. that do not occur apart from the termini) in a specific working polynucleotide are preferred when the use of a corresponding restriction enzyme(s) is not intended to cut the working polynucleotide internally. This allows one to perform restriction digestion reactions to completion or to near completion without incurring unwanted internal cleavage in a working polynucleotide.

According to a preferred aspect, it is thus preferable to use restriction sites that are not contained, or alternatively that are not expected to be contained, or alternatively that are unlikely to be contained (e.g. when sequence information regarding a working polynucleotide is incomplete) internally in a polynucleotide to be subjected to end-selection. In accordance with this aspect, it is appreciated that restriction sites that occur relatively infrequently are usually preferred over those that occur more frequently. On the other hand it is also appreciated that there are occasions where internal cleavage of a polypeptide is desired, e.g. to achieve recombination or other mutagenic procedures along with end-selection.

In accord with this invention, it is also appreciated that methods (e.g. mutagenesis methods) can be used to remove unwanted internal restriction sites. It is also appreciated that a partial digestion reaction (i.e. a digestion reaction that proceeds to partial completion) can be used to achieve digestion at a recognition site in a terminal region while sparing a susceptible restriction site that occurs internally in a polynucleotide and that is recognized by the same enzyme. In one aspect, partial digest are useful because it is appreciated that certain enzymes show preferential cleavage of the same recognition sequence depending on the location and environment in which the recognition sequence occurs. For example, it is appreciated that, while lambda DNA has 5 *EcoR* I sites, cleavage of the site nearest to the right terminus has been reported to occur 10 times faster than the sites in the middle of the molecule. Also, for example, it has been reported that, while *Sac* II has four sites on lambda DNA, the three clustered centrally in lambda are cleaved 50 times faster than the remaining site near the terminus (at nucleotide 40,386). Summarily, site preferences have been reported for various enzymes by many investigators (e.g., Thomas and Davis, 1975; Forsblum et al, 1976; Nath and Azzolina, 1981; Brown and Smith, 1977; Gingeras and Brooks, 1983; Krüger et al, 1988; Conrad and Topal, 1989; Oller et al, 1991; Topal, 1991; and Pein, 1991; to name but a few). It is appreciated that any empirical observations as well as any mechanistic understandings of site preferences by any serviceable polynucleotide-acting enzymes, whether currently available or to be procured in the future, may be serviceable in end-selection according to this invention.

It is also appreciated that protection methods can be used to selectively protect specified restriction sites (e.g. internal sites) against unwanted digestion by enzymes that would otherwise cut a working polypeptide in response to the presence of those sites; and that such protection methods include modifications such as methylations and base substitutions (e.g. U instead of T) that inhibit an unwanted enzyme activity. It is appreciated that there are limited numbers of available restriction enzymes that are rare enough (e.g. having very long recognition sequences) to create large (e.g. megabase-long) restriction fragments, and that protection approaches (e.g. by methylation) are serviceable for increasing the rarity of enzyme cleavage sites. The use of *M.Fnu* II (mCGCG) to increase the apparent rarity of *Not* I approximately twofold is but one example among many (Qiang et al, 1990; Nelson et al, 1984; Maxam and Gilbert, 1980; Raleigh and Wilson, 1986).

According to a preferred aspect of this invention, it is provided that, in general, the use of rare restriction sites is preferred. It is appreciated that, in general, the frequency of occurrence of a restriction site is determined by the number of nucleotides contained therein, as well as by the ambiguity of the base requirements contained therein. Thus, in a non-limiting exemplification, it is appreciated that, in general, a restriction site composed of, for example, 8 specific nucleotides (e.g. the *Not* I site or GC/GGCCGC, with an estimated relative occurrence of 1 in 4^8 , i.e. 1 in 65,536, random 8-mers) is relatively more infrequent than one composed of, for example, 6 nucleotides (e.g. the *Sma* I site or CCC/GGG, having an estimated relative occurrence of 1 in 4^6 , i.e. 1 in 4,096, random 6-mers), which in turn is relatively more infrequent than one composed of, for example, 4 nucleotides (e.g. the *Msp* I site or C/CGG, having an estimated relative occurrence of 1 in 4^4 , i.e. 1 in 256, random 4-mers). Moreover, in another non-limiting exemplification, it is appreciated that, in general, a restriction site having no ambiguous (but only specific) base requirements (e.g. the *Fin* I site or GTCCC, having an estimated relative occurrence of 1 in 4^5 , i.e. 1 in 1024, random 5-mers) is relatively more infrequent than one having an ambiguous W (where W = A or T) base requirement (e.g. the *Ava* II site or G/GWCC, having an estimated relative occurrence of 1 in $4 \times 4 \times 2 \times 4 \times 4$ - i.e. 1 in 512 - random 5-mers), which in turn is relatively more infrequent than one having an ambiguous N (where N = A or C or G or T) base requirement (e.g. the *Asu* I site or G/GNCC, having an estimated relative occurrence of 1 in $4 \times 4 \times 1 \times 4 \times 4$, i.e. 1 in 256 - random 5-mers). These

relative occurrences are considered general estimates for actual polynucleotides, because it is appreciated that specific nucleotide bases (not to mention specific nucleotide sequences) occur with dissimilar frequencies in specific polynucleotides, in specific species of organisms, and in specific groupings of organisms. For example, it is appreciated that the % G+C contents of different species of organisms are often very different and wide ranging.

The use of relatively more infrequent restriction sites as a selection marker include - in a non-limiting fashion - preferably those sites composed at least a 4 nucleotide sequence, more preferably those composed of at least a 5 nucleotide sequence, more preferably still those composed at least a 6 nucleotide sequence (e.g. the *Bam*H I site or G/GATCC, the *Bgl* II site or A/GATCT, the *Pst* I site or CTGCA/G, and the *Xba* I site or T/CTAGA), more preferably still those composed at least a 7 nucleotide sequence, more preferably still those composed of an 8 nucleotide sequence nucleotide sequence (e.g. the *Asc* I site or GG/CGCGCC, the *Not* I site or GC/GGCCGC, the *Pac* I site or TTAAT/TAA, the *Pme* I site or GTTT/AAAC, the *Srf*I site or GCCC/GGGC, the *Sse*838 I site or CCTGCA/GG, and the *Swa* I site or ATTT/AAAT), more preferably still those composed of a 9 nucleotide sequence, and even more preferably still those composed of at least a 10 nucleotide sequence (e.g. the *Bsp*G I site or CG/CGCTGGAC). It is further appreciated that some restriction sites (e.g. for class IIS enzymes) are comprised of a portion of relatively high specificity (i.e. a portion containing a principal determinant of the frequency of occurrence of the restriction site) and a portion of relatively low specificity; and that a site of cleavage may or may not be contained within a portion of relatively low specificity. For example, in the *Eco*57 I site or CTGAAG(16/14), there is a portion of relatively high specificity (i.e. the CTGAAG portion) and a portion of relatively low specificity (i.e. the N16 sequence) that contains a site of cleavage.

In another preferred embodiment of this invention, a serviceable end-selection marker is a terminal sequence that is recognized by a polynucleotide-acting enzyme that recognizes a specific polynucleotide sequence. In a preferred aspect of this invention, serviceable polynucleotide-acting enzymes also include other enzymes in addition to classic type II restriction enzymes. According to this preferred aspect of this invention,

serviceable polynucleotide-acting enzymes also include gyrases, helicases, recombinases, relaxases, and any enzymes related thereto.

Among preferred examples are topoisomerases (which have been categorized by some as a subset of the gyrases) and any other enzymes that have polynucleotide-cleaving activity (including preferably polynucleotide-nicking activity) &/or polynucleotide-ligating activity. Among preferred topoisomerase enzymes are topoisomerase I enzymes, which is available from many commercial sources (Epicentre Technologies, Madison, WI; Invitrogen, Carlsbad, CA; Life Technologies, Gaithersburg, MD) and conceivably even more private sources. It is appreciated that similar enzymes may be developed in the future that are serviceable for end-selection as provided herein. A particularly preferred topoisomerase I enzyme is a topoisomerase I enzyme of vaccinia virus origin, that has a specific recognition sequence (e.g. 5'...AAGGG...3') and has both polynucleotide-nicking activity and polynucleotide-ligating activity. Due to the specific nicking-activity of this enzyme (cleavage of one strand), internal recognition sites are not prone to polynucleotide destruction resulting from the nicking activity (but rather remain annealed) at a temperature that causes denaturation of a terminal site that has been nicked. Thus for use in end-selection, it is preferable that a nicking site for topoisomerase-based end-selection be no more than 100 nucleotides from a terminus, more preferably no more than 50 nucleotides from a terminus, more preferably still no more than 25 nucleotides from a terminus, even more preferably still no more than 20 nucleotides from a terminus, even more preferably still no more than 15 nucleotides from a terminus, even more preferably still no more than 10 nucleotides from a terminus, even more preferably still no more than 8 nucleotides from a terminus, even more preferably still no more than 6 nucleotides from a terminus, and even more preferably still no more than 4 nucleotides from a terminus.

In a particularly preferred exemplification that is non-limiting yet clearly illustrative, it is appreciated that when a nicking site for topoisomerase-based end-selection is 4 nucleotides from a terminus, nicking produces a single stranded oligo of 4 bases (in a terminal region) that can be denatured from its complementary strand in an end-selectable polynucleotide; this provides a sticky end (comprised of 4 bases) in a polynucleotide that is serviceable for an ensuing ligation reaction. To accomplish ligation to a cloning vector (preferably an expression vector), compatible sticky ends can be

generated in a cloning vector by any means including by restriction enzyme-based means. The terminal nucleotides (comprised of 4 terminal bases in this specific example) in an end-selectable polynucleotide terminus are thus wisely chosen to provide compatibility with a sticky end generated in a cloning vector to which the polynucleotide is to be ligated.

On the other hand, internal nicking of an end-selectable polynucleotide, e.g. 500 bases from a terminus, produces a single stranded oligo of 500 bases that is not easily denatured from its complementary strand, but rather is serviceable for repair (e.g. by the same topoisomerase enzyme that produced the nick).

This invention thus provides a method - e.g. that is vaccinia topoisomerase-based &/or type II (or IIS) restriction endonuclease-based &/or type III restriction endonuclease-based &/or nicking enzyme-based (e.g. using N. *Bst*NB I) - for producing a sticky end in a working polynucleotide, which end is ligation compatible, and which end can be comprised of at least a 1 base overhang. Preferably such a sticky end is comprised of at least a 2-base overhang, more preferably such a sticky end is comprised of at least a 3-base overhang, more preferably still such a sticky end is comprised of at least a 4-base overhang, even more preferably still such a sticky end is comprised of at least a 5-base overhang, even more preferably still such a sticky end is comprised of at least a 6-base overhang. Such a sticky end may also be comprised of at least a 7-base overhang, or at least an 8-base overhang, or at least a 9-base overhang, or at least a 10-base overhang, or at least 15-base overhang, or at least a 20-base overhang, or at least a 25-base overhang, or at least a 30-base overhang. These overhangs can be comprised of any bases, including A, C, G, or T.

It is appreciated that sticky end overhangs introduced using topoisomerase or a nicking enzyme (e.g. using N. *Bst*NB I) can be designed to be unique in a ligation environment, so as to prevent unwanted fragment reassemblies, such as self-dimerizations and other unwanted concatamerizations.

According to one aspect of this invention, a plurality of sequences (which may but do not necessarily overlap) can be introduced into a terminal region of an end-selectable polynucleotide by the use of an oligo in a polymerase-based reaction. In a relevant, but by

no means limiting example, such an oligo can be used to provide a preferred 5' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky end (preferably by a vaccinia topoisomerase I), a ribosome binding site (i.e. and "RBS", that is preferably serviceable for expression cloning), and optional linker sequence followed by an ATG start site and a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the polymerase-based reaction). Thus, according to this example, a serviceable oligo (which may be termed a forward primer) can have the sequence: 5'[terminal sequence = (N)₁₋₁₀][topoisomerase I site & RBS = AAGGGAGGAG][linker = (N)₁₋₁₀₀][start codon and template-specific sequence = ATG(N)₀₋₁₀₀]3'.

Analogously, in a relevant, but by no means limiting example, an oligo can be used to provide a preferred 3' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky end (preferably by a vaccinia topoisomerase I), and optional linker sequence followed by a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the polymerase-based reaction). Thus, according to this example, a serviceable oligo (which may be termed a reverse primer) can have the sequence: 5'[terminal sequence = (N)₁₋₁₀][topoisomerase I site = AAGGG][linker = (N)₁₋₁₀₀][template-specific sequence = (N)₀₋₁₀₀]3'.

It is appreciated that, end-selection can be used to distinguish and separate parental template molecules (e.g. to be subjected to mutagenesis) from progeny molecules (e.g. generated by mutagenesis). For example, a first set of primers, lacking in a topoisomerase I recognition site, can be used to modify the terminal regions of the parental molecules (e.g. in polymerase-based amplification). A different second set of primers (e.g. having a topoisomerase I recognition site) can then be used to generate mutated progeny molecules (e.g. using any polynucleotide chimerization method, such as interrupted synthesis, template-switching polymerase-based amplification, or interrupted synthesis; or using saturation mutagenesis; or using any other method for introducing a topoisomerase I recognition site into a mutagenized progeny molecule as disclosed herein) from the amplified template molecules. The use of topoisomerase I-based end-selection can then facilitate, not only discernment, but selective topoisomerase I-based ligation of the desired progeny molecules.

Annealment of a second set of primers to thusly amplified parental molecules can be facilitated by including sequences in a first set of primers (i.e. primers used for amplifying a set parental molecules) that are similar to a topoisomerase I recognition site, yet different enough to prevent functional topoisomerase I enzyme recognition. For example, sequences that diverge from the AAGGG site by anywhere from 1 base to all 5 bases can be incorporated into a first set of primers (to be used for amplifying the parental templates prior to subjection to mutagenesis). In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be amplified using the following exemplary – but by no means limiting – set of forward and reverse primers:

Forward Primer: 5' CTAGAAGAGAGGAGAAAACCATG(N)₁₀₋₁₀₀ 3', and
Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG(N)₁₀₋₁₀₀ 3'

According to this specific example of a first set of primers, (N)₁₀₋₁₀₀ represents preferably a 10 to 100 nucleotide-long template-specific sequence, more preferably a 10 to 50 nucleotide-long template-specific sequence, more preferably still a 10 to 30 nucleotide-long template-specific sequence, and even more preferably still a 15 to 25 nucleotide-long template-specific sequence.

According to a specific, but non-limiting aspect, it is thus provided that, after this amplification (using a disclosed first set of primers lacking in a true topoisomerase I recognition site), amplified parental molecules can then be subjected to mutagenesis using one or more sets of forward and reverse primers that do have a true topoisomerase I recognition site. In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be used as templates for the generation of a mutagenized progeny molecule using the following exemplary – but by no means limiting – second set of forward and reverse primers:

Forward Primer: 5' CTAGAAGGGAGGAGAAAACCATG 3'
Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG 3' (contains *Asc* I recognition sequence)

It is appreciated that any number of different primers sets not specifically mentioned can be used as first, second, or subsequent sets of primers for end-selection consistent with this invention. Notice that type II restriction enzyme sites can be incorporated (e.g. an *Asc* I site in the above example). It is provided that, in addition to the other sequences mentioned, the experimentalist can incorporate one or more N,N,G/T triplets into a serviceable primer in order to subject a working polynucleotide to saturation mutagenesis. Summarily, use of a second and/or subsequent set of primers can achieve dual goals of introducing a topoisomerase I site and of generating mutations in a progeny polynucleotide.

Thus, according to one use provided, a serviceable end-selection marker is an enzyme recognition site that allows an enzyme to cleave (including nick) a polynucleotide at a specified site, to produce a ligation-compatible end upon denaturation of a generated single stranded oligo. Ligation of the produced polynucleotide end can then be accomplished by the same enzyme (e.g. in the case of vaccinia virus topoisomerase I), or alternatively with the use of a different enzyme. According to one aspect of this invention, any serviceable end-selection markers, whether like (e.g. two vaccinia virus topoisomerase I recognition sites) or unlike (e.g. a class II restriction enzyme recognition site and a vaccinia virus topoisomerase I recognition site) can be used in combination to select a polynucleotide. Each selectable polynucleotide can thus have one or more end-selection markers, and they can be like or unlike end-selection markers. In a particular aspect, a plurality of end-selection markers can be located on one end of a polynucleotide and can have overlapping sequences with each other.

It is important to emphasize that any number of enzymes, whether currently in existence or to be developed, can be serviceable in end-selection according to this invention. For example, in a particular aspect of this invention, a nicking enzyme (e.g. N. *Bst* I, which cleaves only one strand at 5'...GAGTCNNNN/N...3') can be used in conjunction with a source of polynucleotide-ligating activity in order to achieve end-selection. According to this embodiment, a recognition site for N. *Bst* I – instead of a recognition site for topoisomerase I – should be incorporated into an end-selectable polynucleotide (whether end-selection is used for selection of a mutagenized progeny molecule or whether end-selection is used apart from any mutagenesis procedure).

It is appreciated that the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation has several advantages over previously available selection methods. In sum, this approach allows one to achieve direction cloning (including expression cloning). Specifically, this approach can be used for the achievement of: direct ligation (i.e. without subsection to a classic restriction-purification-ligation reaction, that is susceptible to a multitude of potential problems from an initial restriction reaction to a ligation reaction dependent on the use of T4 DNA ligase); separation of progeny molecules from original template molecules (e.g. original template molecules lack topoisomerase I sites that not introduced until after mutagenesis), obviation of the need for size separation steps (e.g. by gel chromatography or by other electrophoretic means or by the use of size-exclusion membranes), preservation of internal sequences (even when topoisomerase I sites are present), obviation of concerns about unsuccessful ligation reactions (e.g. dependent on the use of T4 DNA ligase, particularly in the presence of unwanted residual restriction enzyme activity), and facilitated expression cloning (including obviation of frame shift concerns). Concerns about unwanted restriction enzyme-based cleavages – especially at internal restriction sites (or even at often unpredictable sites of unwanted star activity) in a working polynucleotide – that are potential sites of destruction of a working polynucleotide can also be obviated by the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation.

3.3. ADDITIONAL SCREENING METHODS

Peptide Display Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by peptide display methods, wherein an associated polynucleotide encodes a displayed peptide which is screened for a phenotype (e.g., for affinity for a predetermined receptor (ligand)).

An increasingly important aspect of bio-pharmaceutical drug development and molecular biology is the identification of peptide structures, including the primary amino acid sequences, of peptides or peptidomimetics that interact with biological

macromolecules. one method of identifying peptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library or peptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the peptide.

In addition to direct chemical synthesis methods for generating peptide libraries, several recombinant DNA methods also have been reported. One type involves the display of a peptide sequence, antibody, or other protein on the surface of a bacteriophage particle or cell. Generally, in these methods each bacteriophage particle or cell serves as an individual library member displaying a single species of displayed peptide in addition to the natural bacteriophage or cell protein sequences. Each bacteriophage or cell contains the nucleotide sequence information encoding the particular displayed peptide sequence; thus, the displayed peptide sequence can be ascertained by nucleotide sequence determination of an isolated library member.

A well-known peptide display method involves the presentation of a peptide sequence on the surface of a filamentous bacteriophage, typically as a fusion with a bacteriophage coat protein. The bacteriophage library can be incubated with an immobilized, predetermined macromolecule or small molecule (e.g., a receptor) so that bacteriophage particles which present a peptide sequence that binds to the immobilized macromolecule can be differentially partitioned from those that do not present peptide sequences that bind to the predetermined macromolecule. The bacteriophage particles (i.e., library members) which are bound to the immobilized macromolecule are then recovered and replicated to amplify the selected bacteriophage sub-population for a subsequent round of affinity enrichment and phage replication. After several rounds of affinity enrichment and phage replication, the bacteriophage library members that are thus selected are isolated and the nucleotide sequence encoding the displayed peptide sequence is determined, thereby identifying the sequence(s) of peptides that bind to the predetermined macromolecule (e.g., receptor). Such methods are further described in PCT patent publications WO 91/17271, WO 91/18980, WO 91/19818 and WO 93/08278.

The latter PCT publication describes a recombinant DNA method for the display of peptide ligands that involves the production of a library of fusion proteins with each fusion protein composed of a first polypeptide portion, typically comprising a variable sequence, that is available for potential binding to a predetermined macromolecule, and a second polypeptide portion that binds to DNA, such as the DNA vector encoding the individual fusion protein. When transformed host cells are cultured under conditions that allow for expression of the fusion protein, the fusion protein binds to the DNA vector encoding it. Upon lysis of the host cell, the fusion protein/vector DNA complexes can be screened against a predetermined macromolecule in much the same way as bacteriophage particles are screened in the phage-based display system, with the replication and sequencing of the DNA vectors in the selected fusion protein/vector DNA complexes serving as the basis for identification of the selected library peptide sequence(s).

Other systems for generating libraries of peptides and like polymers have aspects of both the recombinant and *in vitro* chemical synthesis methods. In these hybrid methods, cell-free enzymatic machinery is employed to accomplish the *in vitro* synthesis of the library members (i.e., peptides or polynucleotides). In one type of method, RNA molecules with the ability to bind a predetermined protein or a predetermined dye molecule were selected by alternate rounds of selection and PCR amplification (Tuerk and Gold, 1990; Ellington and Szostak, 1990). A similar technique was used to identify DNA sequences which bind a predetermined human transcription factor (Thiesen and Bach, 1990; Beaudry and Joyce, 1992; PCT patent publications WO 92/05258 and WO 92/14843). In a similar fashion, the technique of *in vitro* translation has been used to synthesize proteins of interest and has been proposed as a method for generating large libraries of peptides. These methods which rely upon *in vitro* translation, generally comprising stabilized polysome complexes, are described further in PCT patent publications WO 88/08453, WO 90/05785, WO 90/07003, WO 91/02076, WO 91/05058, and WO 92/02536. Applicants have described methods in which library members comprise a fusion protein having a first polypeptide portion with DNA binding activity and a second polypeptide portion having the library member unique peptide sequence; such methods are suitable for use in cell-free *in vitro* selection formats, among others.

The displayed peptide sequences can be of varying lengths, typically from 3-5000 amino acids long or longer, frequently from 5-100 amino acids long, and often from about 8-15 amino acids long. A library can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Portions or all of the displayed peptide sequence(s) can be random, pseudorandom, defined set kernel, fixed, or the like. The present display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

The present invention also provides random, pseudorandom, and defined sequence framework peptide libraries and methods for generating and screening those libraries to identify useful compounds (e.g., peptides, including single-chain antibodies) that bind to receptor molecules or epitopes of interest or gene products that modify peptides or RNA in a desired fashion. The random, pseudorandom, and defined sequence framework peptides are produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.

A method of affinity enrichment allows a very large library of peptides and single-chain antibodies to be screened and the polynucleotide sequence encoding the desired peptide(s) or single-chain antibodies to be selected. The polynucleotide can then be isolated and shuffled to recombine combinatorially the amino acid sequence of the selected peptide(s) (or predetermined portions thereof) or single-chain antibodies (or just VHI, VLI or CDR portions thereof). Using these methods, one can identify a peptide or single-chain antibody as having a desired binding affinity for a molecule and can exploit the process of shuffling to converge rapidly to a desired high-affinity peptide or scfv. The peptide or antibody can then be synthesized in bulk by conventional means for any suitable use (e.g., as a therapeutic or diagnostic agent).

A significant advantage of the present invention is that no prior information regarding an expected ligand structure is required to isolate peptide ligands or antibodies of interest. The peptide identified can have biological activity, which is meant to include at least specific binding affinity for a selected receptor molecule and, in some instances, will further include the ability to block the binding of other compounds, to stimulate or inhibit metabolic pathways, to act as a signal or messenger, to stimulate or inhibit cellular activity, and the like.

The present invention also provides a method for shuffling a pool of polynucleotide sequences selected by affinity screening a library of polysomes displaying nascent peptides (including single-chain antibodies) for library members which bind to a predetermined receptor (e.g., a mammalian proteinaceous receptor such as, for example, a peptidergic hormone receptor, a cell surface receptor, an intracellular protein which binds to other protein(s) to form intracellular protein complexes such as hetero-dimers and the like) or epitope (e.g., an immobilized protein, glycoprotein, oligosaccharide, and the like).

Polynucleotide sequences selected in a first selection round (typically by affinity selection for binding to a receptor (e.g., a ligand)) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral sequences (i.e., having insubstantial functional effect on binding), such as for example by back-crossing with a wild-type or naturally-occurring sequence substantially identical to a selected sequence to produce native-like functional peptides, which may be less immunogenic. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined receptor (ligand).

Prior to or concomitant with the shuffling of selected sequences, the sequences can be mutagenized. In one embodiment, selected library members are cloned in a prokaryotic

vector (e.g., plasmid, phagemid, or bacteriophage) wherein a collection of individual colonies (or plaques) representing discrete library members are produced. Individual selected library members can then be manipulated (e.g., by site-directed mutagenesis, cassette mutagenesis, chemical mutagenesis, PCR mutagenesis, and the like) to generate a collection of library members representing a kernel of sequence diversity based on the sequence of the selected library member. The sequence of an individual selected library member or pool can be manipulated to incorporate random mutation, pseudorandom mutation, defined kernel mutation (i.e., comprising variant and invariant residue positions and/or comprising variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), codon-based mutation, and the like, either segmentally or over the entire length of the individual selected library member sequence. The mutagenized selected library members are then shuffled by *in vitro* and/or *in vivo* recombinatorial shuffling as disclosed herein.

The invention also provides peptide libraries comprising a plurality of individual library members of the invention, wherein (1) each individual library member of said plurality comprises a sequence produced by shuffling of a pool of selected sequences, and (2) each individual library member comprises a variable peptide segment sequence or single-chain antibody segment sequence which is distinct from the variable peptide segment sequences or single-chain antibody sequences of other individual library members in said plurality (although some library members may be present in more than one copy per library due to uneven amplification, stochastic probability, or the like).

The invention also provides a product-by-process, wherein selected polynucleotide sequences having (or encoding a peptide having) a predetermined binding specificity are formed by the process of: (1) screening a displayed peptide or displayed single-chain antibody library against a predetermined receptor (e.g., ligand) or epitope (e.g., antigen macromolecule) and identifying and/or enriching library members which bind to the predetermined receptor or epitope to produce a pool of selected library members, (2) shuffling by recombination the selected library members (or amplified or cloned copies thereof) which binds the predetermined epitope and has been thereby isolated and/or enriched from the library to generate a shuffled library, and (3) screening the shuffled library against the predetermined receptor (e.g., ligand) or epitope (e.g., antigen

macromolecule) and identifying and/or enriching shuffled library members which bind to the predetermined receptor or epitope to produce a pool of selected shuffled library members.

Antibody Display and Screening Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by antibody display methods, wherein an associated polynucleotide encodes a displayed antibody which is screened for a phenotype (e.g., for affinity for binding a predetermined antigen (ligand)).

Various molecular genetic approaches have been devised to capture the vast immunological repertoire represented by the extremely large number of distinct variable regions which can be present in immunoglobulin chains. The naturally-occurring germ line immunoglobulin heavy chain locus is composed of separate tandem arrays of variable segment genes located upstream of a tandem array of diversity segment genes, which are themselves located upstream of a tandem array of joining (i) region genes, which are located upstream of the constant region genes. During B lymphocyte development, V-D-J rearrangement occurs wherein a heavy chain variable region gene (VH) is formed by rearrangement to form a fused D segment followed by rearrangement with a V segment to form a V-D-J joined product gene which, if productively rearranged, encodes a functional variable region (VH) of a heavy chain. Similarly, light chain loci rearrange one of several V segments with one of several J segments to form a gene encoding the variable region (VL) of a light chain.

The vast repertoire of variable regions possible in immunoglobulins derives in part from the numerous combinatorial possibilities of joining V and i segments (and, in the case of heavy chain loci, D segments) during rearrangement in B cell development. Additional sequence diversity in the heavy chain variable regions arises from non-uniform rearrangements of the D segments during V-D-J joining and from N region addition. Further, antigen-selection of specific B cell clones selects for higher affinity variants having non-germline mutations in one or both of the heavy and light chain variable regions; a phenomenon referred to as "affinity maturation" or "affinity sharpening".

Typically, these "affinity sharpening" mutations cluster in specific areas of the variable region, most commonly in the complementarity-determining regions (CDRs).

In order to overcome many of the limitations in producing and identifying high-affinity immunoglobulins through antigen-stimulated B cell development (i.e., immunization), various prokaryotic expression systems have been developed that can be manipulated to produce combinatorial antibody libraries which may be screened for high-affinity antibodies to specific antigens. Recent advances in the expression of antibodies in *Escherichia coli* and bacteriophage systems (see "alternative peptide display methods", *infra*) have raised the possibility that virtually any specificity can be obtained by either cloning antibody genes from characterized hybridomas or by de novo selection using antibody gene libraries (e.g., from Ig cDNA).

Combinatorial libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened as bacteriophage plaques or as colonies of lysogens (Huse et al, 1989); Caton and Koprowski, 1990; Mullinax et al, 1990; Persson et al, 1991). Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (Kang et al, 1991; Clackson et al, 1991; McCafferty et al, 1990; Burton et al, 1991; Hoogenboom et al, 1991; Chang et al, 1991; Breitling et al, 1991; Marks et al, 1991, p. 581; Barbas et al, 1992; Hawkins and Winter, 1992; Marks et al, 1992, p. 779; Marks et al, 1992, p. 16007; and Lowman et al, 1991; Lerner et al, 1992; all incorporated herein by reference). Typically, a bacteriophage antibody display library is screened with a receptor (e.g., polypeptide, carbohydrate, glycoprotein, nucleic acid) that is immobilized (e.g., by covalent linkage to a chromatography resin to enrich for reactive phage by affinity chromatography) and/or labeled (e.g., to screen plaque or colony lifts).

One particularly advantageous approach has been the use of so-called single-chain fragment variable (scfv) libraries (Marks et al, 1992, p. 779; Winter and Milstein, 1991; Clackson et al, 1991; Marks et al, 1991, p. 581; Chaudhary et al, 1990; Chiswell et al, 1992; McCafferty et al, 1990; and Huston et al, 1988). Various embodiments of scfv libraries displayed on bacteriophage coat proteins have been described.

Beginning in 1988, single-chain analogues of Fv fragments and their fusion proteins have been reliably generated by antibody engineering methods. The first step generally involves obtaining the genes encoding VH and VL domains with desired binding properties; these V genes may be isolated from a specific hybridoma cell line, selected from a combinatorial V-gene library, or made by V gene synthesis. The single-chain Fv is formed by connecting the component V genes with an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)₃ or equivalent linker peptide(s). The linker bridges the C-terminus of the first V region and N-terminus of the second, ordered as either VH-linker-VL or VL-linker-VH. In principle, the scfv binding site can faithfully replicate both the affinity and specificity of its parent antibody combining site.

Thus, scfv fragments are comprised of VH and VL domains linked into a single polypeptide chain by a flexible linker peptide. After the scfv genes are assembled, they are cloned into a phagemid and expressed at the tip of the M13 phage (or similar filamentous bacteriophage) as fusion proteins with the bacteriophage PIII (gene 3) coat protein. Enriching for phage expressing an antibody of interest is accomplished by panning the recombinant phage displaying a population scfv for binding to a predetermined epitope (e.g., target antigen, receptor).

The linked polynucleotide of a library member provides the basis for replication of the library member after a screening or selection procedure, and also provides the basis for the determination, by nucleotide sequencing, of the identity of the displayed peptide sequence or VH and VL amino acid sequence. The displayed peptide (s) or single-chain antibody (e.g., scfv) and/or its VH and VL domains or their CDRs can be cloned and expressed in a suitable expression system. Often polynucleotides encoding the isolated VH and VL domains will be ligated to polynucleotides encoding constant regions (CH and CL) to form polynucleotides encoding complete antibodies (e.g., chimeric or fully-human), antibody fragments, and the like. Often polynucleotides encoding the isolated CDRs will be grafted into polynucleotides encoding a suitable variable region framework (and optionally constant regions) to form polynucleotides encoding complete antibodies (e.g., humanized or fully-human), antibody fragments, and the like. Antibodies can be used to isolate preparative quantities of the antigen by immunoaffinity

chromatography. Various other uses of such antibodies are to diagnose and/or stage disease (e.g., neoplasia) and for therapeutic application to treat disease, such as for example: neoplasia, autoimmune disease, AIDS, cardiovascular disease, infections, and the like.

Various methods have been reported for increasing the combinatorial diversity of a scfv library to broaden the repertoire of binding species (idiotype spectrum). The use of PCR has permitted the variable regions to be rapidly cloned either from a specific hybridoma source or as a gene library from non-immunized cells, affording combinatorial diversity in the assortment of VH and VL cassettes which can be combined. Furthermore, the VH and VL cassettes can themselves be diversified, such as by random, pseudorandom, or directed mutagenesis. Typically, VH and VL cassettes are diversified in or near the complementarity-determining regions (CDRS), often the third CDR, CDR3. Enzymatic inverse PCR mutagenesis has been shown to be a simple and reliable method for constructing relatively large libraries of scfv site-directed hybrids (Stemmer et al, 1993), as has error-prone PCR and chemical mutagenesis (Deng et al, 1994). Riechmann (Riechmann et al, 1993) showed semi-rational design of an antibody scfv fragment using site-directed randomization by degenerate oligonucleotide PCR and subsequent phage display of the resultant scfv hybrids. Barbas (Barbas et al, 1992) attempted to circumvent the problem of limited repertoire sizes resulting from using biased variable region sequences by randomizing the sequence in a synthetic CDR region of a human tetanus toxoid-binding Fab.

CDR randomization has the potential to create approximately 1×10^{20} CDRs for the heavy chain CDR3 alone, and a roughly similar number of variants of the heavy chain CDR1 and CDR2, and light chain CDR1-3 variants. Taken individually or together, the combination possibilities of CDR randomization of heavy and/or light chains requires generating a prohibitive number of bacteriophage clones to produce a clone library representing all possible combinations, the vast majority of which will be non-binding. Generation of such large numbers of primary transformants is not feasible with current transformation technology and bacteriophage display systems. For example, Barbas (Barbas et al, 1992) only generated 5×10^7 transformants, which represents only a tiny fraction of the potential diversity of a library of thoroughly randomized CDRS.

Despite these substantial limitations, bacteriophage display of scfv have already yielded a variety of useful antibodies and antibody fusion proteins. A bispecific single chain antibody has been shown to mediate efficient tumor cell lysis (Gruber et al, 1994). Intracellular expression of an anti-Rev scfv has been shown to inhibit HIV-1 virus replication *in vitro* (Duan et al, 1994), and intracellular expression of an anti-p21rar, scfv has been shown to inhibit meiotic maturation of *Xenopus* oocytes (Biocca et al, 1993). Recombinant scfv which can be used to diagnose HIV infection have also been reported, demonstrating the diagnostic utility of scfv (Lilley et al, 1994). Fusion proteins wherein an scFv is linked to a second polypeptide, such as a toxin or fibrinolytic activator protein, have also been reported (Holvost et al, 1992; Nicholls et al, 1993).

If it were possible to generate scfv libraries having broader antibody diversity and overcoming many of the limitations of conventional CDR mutagenesis and randomization methods which can cover only a very tiny fraction of the potential sequence combinations, the number and quality of scfv antibodies suitable for therapeutic and diagnostic use could be vastly improved. To address this, the *in vitro* and *in vivo* shuffling methods of the invention are used to recombine CDRs which have been obtained (typically via PCR amplification or cloning) from nucleic acids obtained from selected displayed antibodies. Such displayed antibodies can be displayed on cells, on bacteriophage particles, on polysomes, or any suitable antibody display system wherein the antibody is associated with its encoding nucleic acid(s). In a variation, the CDRs are initially obtained from mRNA (or cDNA) from antibody-producing cells (e.g., plasma cells/splenocytes from an immunized wild-type mouse, a human, or a transgenic mouse capable of making a human antibody as in WO 92/03918, WO 93/12227, and WO 94/25585), including hybridomas derived therefrom.

Polynucleotide sequences selected in a first selection round (typically by affinity selection for displayed antibody binding to an antigen (e.g., a ligand) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination, especially shuffling of CDRs (typically shuffling heavy chain CDRs with other heavy chain CDRs and light chain CDRs with other light chain CDRs) to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences.

The recombined selected polynucleotide sequences are expressed in a selection format as a displayed antibody and subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection until an antibody of the desired binding affinity is obtained. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral antibody framework sequences (i.e., having insubstantial functional effect on antigen binding), such as for example by back-crossing with a human variable region framework to produce human-like sequence antibodies. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined antigen.

Alternatively, or in combination with the noted variations, the valency of the target epitope may be varied to control the average binding affinity of selected scfv library members. The target epitope can be bound to a surface or substrate at varying densities, such as by including a competitor epitope, by dilution, or by other method known to those in the art. A high density (valency) of predetermined epitope can be used to enrich for scfv library members which have relatively low affinity, whereas a low density (valency) can preferentially enrich for higher affinity scfv library members.

For generating diverse variable segments, a collection of synthetic oligonucleotides encoding random, pseudorandom, or a defined sequence kernel set of peptide sequences can be inserted by ligation into a predetermined site (e.g., a CDR). Similarly, the sequence diversity of one or more CDRs of the single-chain antibody cassette(s) can be expanded by mutating the CDR(s) with site-directed mutagenesis, CDR-replacement, and the like. The resultant DNA molecules can be propagated in a host for cloning and amplification prior to shuffling, or can be used directly (i.e., may avoid loss of diversity which may occur upon propagation in a host cell) and the selected library members subsequently shuffled.

Displayed peptide/polynucleotide complexes (library members) which encode a variable segment peptide sequence of interest or a single-chain antibody of interest are selected from the library by an affinity enrichment technique. This is accomplished by means of a immobilized macromolecule or epitope specific for the peptide sequence of

interest, such as a receptor, other macromolecule, or other epitope species. Repeating the affinity selection procedure provides an enrichment of library members encoding the desired sequences, which may then be isolated for pooling and shuffling, for sequencing, and/or for further propagation and affinity enrichment.

The library members without the desired specificity are removed by washing. The degree and stringency of washing required will be determined for each peptide sequence or single-chain antibody of interest and the immobilized predetermined macromolecule or epitope. A certain degree of control can be exerted over the binding characteristics of the nascent peptide/DNA complexes recovered by adjusting the conditions of the binding incubation and the subsequent washing. The temperature, pH, ionic strength, divalent cations concentration, and the volume and duration of the washing will select for nascent peptide/DNA complexes within particular ranges of affinity for the immobilized macromolecule. Selection based on slow dissociation rate, which is usually predictive of high affinity, is often the most practical route. This may be done either by continued incubation in the presence of a saturating amount of free predetermined macromolecule, or by increasing the volume, number, and length of the washes. In each case, the rebinding of dissociated nascent peptide/DNA or peptide/RNA complex is prevented, and with increasing time, nascent peptide/DNA or peptide/RNA complexes of higher and higher affinity are recovered.

Additional modifications of the binding and washing procedures may be applied to find peptides with special characteristics. The affinities of some peptides are dependent on ionic strength or cation concentration. This is a useful characteristic for peptides that will be used in affinity purification of various proteins when gentle conditions for removing the protein from the peptides are required.

One variation involves the use of multiple binding targets (multiple epitope species, multiple receptor species), such that a scfv library can be simultaneously screened for a multiplicity of scfv which have different binding specificities. Given that the size of a scfv library often limits the diversity of potential scfv sequences, it is typically desirable to use scfv libraries of as large a size as possible. The time and economic considerations of generating a number of very large polysome scFv-display libraries can become

prohibitive. To avoid this substantial problem, multiple predetermined epitope species (receptor species) can be concomitantly screened in a single library, or sequential screening against a number of epitope species can be used. In one variation, multiple target epitope species, each encoded on a separate bead (or subset of beads), can be mixed and incubated with a polysome-display scfv library under suitable binding conditions. The collection of beads, comprising multiple epitope species, can then be used to isolate, by affinity selection, scfv library members. Generally, subsequent affinity screening rounds can include the same mixture of beads, subsets thereof, or beads containing only one or two individual epitope species. This approach affords efficient screening, and is compatible with laboratory automation, batch processing, and high throughput screening methods.

A variety of techniques can be used in the present invention to diversify a peptide library or single-chain antibody library, or to diversify, prior to or concomitant with shuffling, around variable segment peptides found in early rounds of panning to have sufficient binding activity to the predetermined macromolecule or epitope. In one approach, the positive selected peptide/polynucleotide complexes (those identified in an early round of affinity enrichment) are sequenced to determine the identity of the active peptides. Oligonucleotides are then synthesized based on these active peptide sequences, employing a low level of all bases incorporated at each step to produce slight variations of the primary oligonucleotide sequences. This mixture of (slightly) degenerate oligonucleotides is then cloned into the variable segment sequences at the appropriate locations. This method produces systematic, controlled variations of the starting peptide sequences, which can then be shuffled. It requires, however, that individual positive nascent peptide/polynucleotide complexes be sequenced before mutagenesis, and thus is useful for expanding the diversity of small numbers of recovered complexes and selecting variants having higher binding affinity and/or higher binding specificity. In a variation, mutagenic PCR amplification of positive selected peptide/polynucleotide complexes (especially of the variable region sequences, the amplification products of which are shuffled *in vitro* and/or *in vivo* and one or more additional rounds of screening is done prior to sequencing. The same general approach can be employed with single-chain antibodies in order to expand the diversity and enhance the binding affinity/specificity, typically by diversifying CDRs or adjacent framework regions prior to or concomitant

with shuffling. If desired, shuffling reactions can be spiked with mutagenic oligonucleotides capable of *in vitro* recombination with the selected library members can be included. Thus, mixtures of synthetic oligonucleotides and PCR produced polynucleotides (synthesized by error-prone or high-fidelity methods) can be added to the *in vitro* shuffling mix and be incorporated into resulting shuffled library members (shufflants).

The present invention of shuffling enables the generation of a vast library of CDR-variant single-chain antibodies. One way to generate such antibodies is to insert synthetic CDRs into the single-chain antibody and/or CDR randomization prior to or concomitant with shuffling. The sequences of the synthetic CDR cassettes are selected by referring to known sequence data of human CDR and are selected in the discretion of the practitioner according to the following guidelines: synthetic CDRs will have at least 40 percent positional sequence identity to known CDR sequences, and preferably will have at least 50 to 70 percent positional sequence identity to known CDR sequences. For example, a collection of synthetic CDR sequences can be generated by synthesizing a collection of oligonucleotide sequences on the basis of naturally-occurring human CDR sequences listed in Kabat (Kabat et al, 1991); the pool (s) of synthetic CDR sequences are calculated to encode CDR peptide sequences having at least 40 percent sequence identity to at least one known naturally-occurring human CDR sequence. Alternatively, a collection of naturally-occurring CDR sequences may be compared to generate consensus sequences so that amino acids used at a residue position frequently (i.e., in at least 5 percent of known CDR sequences) are incorporated into the synthetic CDRs at the corresponding position(s). Typically, several (e.g., 3 to about 50) known CDR sequences are compared and observed natural sequence variations between the known CDRs are tabulated, and a collection of oligonucleotides encoding CDR peptide sequences encompassing all or most permutations of the observed natural sequence variations is synthesized. For example but not for limitation, if a collection of human VH CDR sequences have carboxy-terminal amino acids which are either Tyr, Val, Phe, or Asp, then the pool(s) of synthetic CDR oligonucleotide sequences are designed to allow the carboxy-terminal CDR residue to be any of these amino acids. In some embodiments, residues other than those which naturally-occur at a residue position in the collection of CDR sequences are incorporated: conservative amino acid substitutions are frequently

incorporated and up to 5 residue positions may be varied to incorporate non-conservative amino acid substitutions as compared to known naturally-occurring CDR sequences. Such CDR sequences can be used in primary library members (prior to first round screening) and/or can be used to spike *in vitro* shuffling reactions of selected library member sequences. Construction of such pools of defined and/or degenerate sequences will be readily accomplished by those of ordinary skill in the art.

The collection of synthetic CDR sequences comprises at least one member that is not known to be a naturally-occurring CDR sequence. It is within the discretion of the practitioner to include or not include a portion of random or pseudorandom sequence corresponding to N region addition in the heavy chain CDR; the N region sequence ranges from 1 nucleotide to about 4 nucleotides occurring at V-D and D-J junctions. A collection of synthetic heavy chain CDR sequences comprises at least about 100 unique CDR sequences, typically at least about 1,000 unique CDR sequences, preferably at least about 10,000 unique CDR sequences, frequently more than 50,000 unique CDR sequences; however, usually not more than about 1×10^6 unique CDR sequences are included in the collection, although occasionally 1×10^7 to 1×10^8 unique CDR sequences are present, especially if conservative amino acid substitutions are permitted at positions where the conservative amino acid substituent is not present or is rare (i.e., less than 0.1 percent) in that position in naturally-occurring human CDRS. In general, the number of unique CDR sequences included in a library should not exceed the expected number of primary transformants in the library by more than a factor of 10. Such single-chain antibodies generally bind of about at least 1×10^{-6} M, preferably with an affinity of about at least 5×10^{-7} M, more preferably with an affinity of at least 1×10^{-8} M to 1×10^{-9} M or more, sometimes up to 1×10^{-10} M or more. Frequently, the predetermined antigen is a human protein, such as for example a human cell surface antigen (e. g., CD4, CD8, IL-2 receptor, EGF receptor, PDGF receptor), other human biological macromolecule (e.g., thrombomodulin, protein C, carbohydrate antigen, sialyl Lewis antigen, Lselectin), or nonhuman disease associated macromolecule (e.g., bacterial LPS, virion capsid protein or envelope glycoprotein) and the like.

High affinity single-chain antibodies of the desired specificity can be engineered and expressed in a variety of systems. For example, scfv have been produced in plants

(Firek et al, 1993) and can be readily made in prokaryotic systems (Owens and Young, 1994; Johnson and Bird, 1991). Furthermore, the single-chain antibodies can be used as a basis for constructing whole antibodies or various fragments thereof (Kettleborough et al, 1994). The variable region encoding sequence may be isolated (e.g., by PCR amplification or subcloning) and spliced to a sequence encoding a desired human constant region to encode a human sequence antibody more suitable for human therapeutic uses where immunogenicity is preferably minimized. The polynucleotide(s) having the resultant fully human encoding sequence(s) can be expressed in a host cell (e.g., from an expression vector in a mammalian cell) and purified for pharmaceutical formulation.

The DNA expression constructs will typically include an expression control DNA sequence operably linked to the coding sequences, including naturally-associated or heterologous promoter regions. Preferably, the expression control sequences will be eukaryotic promoter systems in vectors capable of transforming or transfecting eukaryotic host cells. Once the vector has been incorporated into the appropriate host, the host is maintained under conditions suitable for high level expression of the nucleotide sequences, and the collection and purification of the mutant "engineered" antibodies.

As stated previously, the DNA sequences will be expressed in hosts after the sequences have been operably linked to an expression control sequence (i.e., positioned to ensure the transcription and translation of the structural gene). These expression vectors are typically replicable in the host organisms either as episomes or as an integral part of the host chromosomal DNA. Commonly, expression vectors will contain selection markers, e.g., tetracycline or neomycin, to permit detection of those cells transformed with the desired DNA sequences (see, e.g., USPN 4,704,362, which is incorporated herein by reference).

In addition to eukaryotic microorganisms such as yeast, mammalian tissue cell culture may also be used to produce the polypeptides of the present invention (see Winnacker, 1987), which is incorporated herein by reference). Eukaryotic cells are actually preferred, because a number of suitable host cell lines capable of secreting intact immunoglobulins have been developed in the art, and include the CHO cell lines, various COS cell lines, HeLa cells, and myeloma cell lines, but preferably transformed Bcells or

hybridomas. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter, an enhancer (Queen et al, 1986), and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites, and transcriptional terminator sequences. Preferred expression control sequences are promoters derived from immunoglobulin genes, cytomegalovirus, SV40, Adenovirus, Bovine Papilloma Virus, and the like.

Eukaryotic DNA transcription can be increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting sequences of between 10 to 300 bp that increase transcription by a promoter. Enhancers can effectively increase transcription when either 5' or 3' to the transcription unit. They are also effective if located within an intron or within the coding sequence itself. Typically, viral enhancers are used, including SV40 enhancers, cytomegalovirus enhancers, polyoma enhancers, and adenovirus enhancers. Enhancer sequences from mammalian systems are also commonly used, such as the mouse immunoglobulin heavy chain enhancer.

Mammalian expression vector systems will also typically include a selectable marker gene. Examples of suitable markers include, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance. The first two marker genes prefer the use of mutant cell lines that lack the ability to grow without the addition of thymidine to the growth medium. Transformed cells can then be identified by their ability to grow on non-supplemented media. Examples of prokaryotic drug resistance genes useful as markers include genes conferring resistance to G418, mycophenolic acid and hygromycin.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, calcium chloride transfection is commonly utilized for prokaryotic cells, whereas calcium phosphate treatment, lipofection, or electroporation may be used for other cellular hosts. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, liposomes, electroporation, and micro-injection (see, generally, Sambrook et al, 1982 and 1989).

Once expressed, the antibodies, individual mutated immunoglobulin chains, mutated antibody fragments, and other immunoglobulin polypeptides of the invention can be purified according to standard procedures of the art, including ammonium sulfate precipitation, fraction column chromatography, gel electrophoresis and the like (see, generally, Scopes, 1982). Once purified, partially or to homogeneity as desired, the polypeptides may then be used therapeutically or in developing and performing assay procedures, immunofluorescent stainings, and the like (see, generally, Lefkovits and Pernis, 1979 and 1981; Lefkovits, 1997).

The antibodies generated by the method of the present invention can be used for diagnosis and therapy. By way of illustration and not limitation, they can be used to treat cancer, autoimmune diseases, or viral infections. For treatment of cancer, the antibodies will typically bind to an antigen expressed preferentially on cancer cells, such as erbB-2, CEA, CD33, and many other antigens and binding members well known to those skilled in the art.

Two-Hybrid Based Screening Assays

Shuffling can also be used to recombinatorially diversify a pool of selected library members obtained by screening a two-hybrid screening system to identify library members which bind a predetermined polypeptide sequence. The selected library members are pooled and shuffled by *in vitro* and/or *in vivo* recombination. The shuffled pool can then be screened in a yeast two hybrid system to select library members which bind said predetermined polypeptide sequence (e. g., and SH2 domain) or which bind an alternate predetermined polypeptide sequence (e.g., an SH2 domain from another protein species).

An approach to identifying polypeptide sequences which bind to a predetermined polypeptide sequence has been to use a so-called "two-hybrid" system wherein the predetermined polypeptide sequence is present in a fusion protein (Chien et al, 1991). This approach identifies protein-protein interactions *in vivo* through reconstitution of a transcriptional activator (Fields and Song, 1989), the yeast Gal4 transcription protein. Typically, the method is based on the properties of the yeast Gal4 protein, which consists of separable domains responsible for DNA-binding and transcriptional activation.

Polynucleotides encoding two hybrid proteins, one consisting of the yeast Gal4 DNA-binding domain fused to a polypeptide sequence of a known protein and the other consisting of the Gal4 activation domain fused to a polypeptide sequence of a second protein, are constructed and introduced into a yeast host cell. Intermolecular binding between the two fusion proteins reconstitutes the Gal4 DNA-binding domain with the Gal4 activation domain, which leads to the transcriptional activation of a reporter gene (e.g., *lacZ*, *HIS3*) which is operably linked to a Gal4 binding site. Typically, the two-hybrid method is used to identify novel polypeptide sequences which interact with a known protein (Silver and Hunt, 1993; Durfee et al, 1993; Yang et al, 1992; Luban et al, 1993; Hardy et al, 1992; Bartel et al, 1993; and Vojtek et al, 1993). However, variations of the two-hybrid method have been used to identify mutations of a known protein that affect its binding to a second known protein (Li and Fields, 1993; Lalo et al, 1993; Jackson et al, 1993; and Madura et al, 1993). Two-hybrid systems have also been used to identify interacting structural domains of two known proteins (Bardwell et al, 1993; Chakrabarty et al, 1992; Staudinger et al, 1993; and Milne and Weaver 1993) or domains responsible for oligomerization of a single protein (Iwabuchi et al, 1993; Bogerd et al, 1993). Variations of two-hybrid systems have been used to study the *in vivo* activity of a proteolytic enzyme (Dasmahapatra et al, 1992). Alternatively, an *E. coli*/BCCP interactive screening system (Germino et al, 1993; Guarente, 1993) can be used to identify interacting protein sequences (i.e., protein sequences which heterodimerize or form higher order heteromultimers). Sequences selected by a two-hybrid system can be pooled and shuffled and introduced into a two-hybrid system for one or more subsequent rounds of screening to identify polypeptide sequences which bind to the hybrid containing the predetermined binding sequence. The sequences thus identified can be compared to identify consensus sequence(s) and consensus sequence kernels.

In general, standard techniques of recombination DNA technology are described in various publications (e.g. Sambrook et al, 1989; Ausubel et al, 1987; and Berger and Kimmel, 1987) each of which is incorporated herein in its entirety by reference. Polynucleotide modifying enzymes were used according to the manufacturer's recommendations. Oligonucleotides were synthesized on an Applied Biosystems Inc. Model 394 DNA synthesizer using ABI chemicals. If desired, PCR amplimers for

amplifying a predetermined DNA sequence may be selected at the discretion of the practitioner.

One microgram samples of template DNA are obtained and treated with U.V. light to cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template DNA sample. Multiple samples are treated with U.V. light for varying periods of time to obtain template DNA samples with varying numbers of dimers from U.V. exposure.

A random priming kit which utilizes a non-proofreading polymease (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

The present invention is further directed to a method for generating a selected mutant polynucleotide sequence (or a population of selected polynucleotide sequences) typically in the form of amplified and/or cloned polynucleotides, whereby the selected polynucleotide sequences(s) possess at least one desired phenotypic characteristic (e.g., encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, and the like) which can be selected for. One method for identifying hybrid polypeptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library of polypeptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the polypeptide.

In one embodiment, the present invention provides a method for generating libraries of displayed polypeptides or displayed antibodies suitable for affinity interaction screening or phenotypic screening. The method comprises (1) obtaining a first plurality of

selected library members comprising a displayed polypeptide or displayed antibody and an associated polynucleotide encoding said displayed polypeptide or displayed antibody, and obtaining said associated polynucleotides or copies thereof wherein said associated polynucleotides comprise a region of substantially identical sequences, optimally introducing mutations into said polynucleotides or copies, (2) pooling the polynucleotides or copies, (3) producing smaller or shorter polynucleotides by interrupting a random or particularized priming and synthesis process or an amplification process, and (4) performing amplification, preferably PCR amplification, and optionally mutagenesis to homologously recombine the newly synthesized polynucleotides.

It is a particularly preferred object of the invention to provide a process for producing hybrid polynucleotides which express a useful hybrid polypeptide by a series of steps comprising:

- (a) producing polynucleotides by interrupting a polynucleotide amplification or synthesis process with a means for blocking or interrupting the amplification or synthesis process and thus providing a plurality of smaller or shorter polynucleotides due to the replication of the polynucleotide being in various stages of completion;
- (b) adding to the resultant population of single- or double-stranded polynucleotides one or more single- or double-stranded oligonucleotides, wherein said added oligonucleotides comprise an area of identity in an area of heterology to one or more of the single- or double-stranded polynucleotides of the population;
- (c) denaturing the resulting single- or double-stranded oligonucleotides to produce a mixture of single-stranded polynucleotides, optionally separating the shorter or smaller polynucleotides into pools of polynucleotides having various lengths and further optionally subjecting said polynucleotides to a PCR procedure to amplify one or more oligonucleotides comprised by at least one of said polynucleotide pools;
- (d) incubating a plurality of said polynucleotides or at least one pool of said polynucleotides with a polymerase under conditions which result in annealing of said single-stranded polynucleotides at regions of identity between the single-stranded polynucleotides and thus forming of a mutagenized double-stranded polynucleotide chain;
- (e) optionally repeating steps (c) and (d);
- (f) expressing at least one hybrid polypeptide from said polynucleotide chain, or chains; and

(g) screening said at least one hybrid polypeptide for a useful activity.

In a preferred aspect of the invention, the means for blocking or interrupting the amplification or synthesis process is by utilization of UV light, DNA adducts, DNA binding proteins.

In one embodiment of the invention, the DNA adducts, or polynucleotides comprising the DNA adducts, are removed from the polynucleotides or polynucleotide pool, such as by a process including heating the solution comprising the DNA fragments prior to further processing.

Having thus disclosed exemplary embodiments of the present invention, it should be noted by those skilled in the art that the disclosures are exemplary only and that various other alternatives, adaptations and modifications may be made within the scope of the present invention. Accordingly, the present invention is not limited to the specific embodiments as illustrated herein.

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following examples are to be considered illustrative and thus are not limiting of the remainder of the disclosure in any way whatsoever.

4. Transposons

4.1. GENERAL APPLICATIONS

In one aspect, the present invention relates generally to the field of transposable nucleic acid and, more particularly to production and use of a modified transposase enzyme in a system for introducing genetic changes to nucleic acid.

The present invention relates to transposable elements isolated from maize and a process for using the same to identify and isolate genes and to insert desired gene sequences into plants in a heritable manner.

4.2. SPECIFIC METHODOLOGIES

4.2.1. Description Of Transposable Elements

Transposable genetic elements are DNA sequences, found in a wide variety of prokaryotic and eukaryotic organisms, that can move or transpose from one position to another position in a genome. In vivo, intra-chromosomal transpositions as well as transpositions between chromosomal and non-chromosomal genetic material are known. In several systems, transposition is known to be under the control of a transposase enzyme that is typically encoded by the transposable element. The genetic structures and transposition mechanisms of various transposable elements are summarized, for example, in "Transposable Genetic Elements" in "The Encyclopedia of Molecular Biology," Kendrew and Lawrence, Eds., Blackwell Science, Ltd., Oxford (1994), incorporated herein by reference.

Transposable elements (hereinafter "transposons") are natural gene transfer vectors in bacteria, yeast, *Drosophila melanogaster* and other organisms. The best documented examples of transposons in bacteria are those carrying genes that confer antibiotic resistance on the bacterium in which they reside. These transposons tend to accumulate and become a part of bacterial plasmids. The biological properties of the plasmids permit the spread of the plasmids and their passengers, e.g., drug resistance transposons, in bacterial populations.

Transposons have also been used to identify and isolate otherwise inaccessible genes (See Bingham, P. M., Kidwell, M. G. and Rubin, G. M., Cell 29:995-1004 (1982)). That is, the White locus of *Drosophila melanogaster* has been isolated by virtue of the existence of an insertion mutation at the locus caused by a transposon that has been isolated and studied using recombinant DNA technology. Such applications are receiving increasing attention in plants and animals.

The use of transposable elements as deliberate gene transfer vectors evolved from work in bacteria and yeast and, as stated above, has recently been developed into a useful research tool in *Drosophila melanogaster* (See Rubin, G. M. and Spradling, A. C., Science 218:348-353 (1982)). The basic principle on which such applications are based is that transposons are compact genetic units that contain within their sequences essentially all of

the coding information required for transposition. Although the transposition functions are only now beginning to be identified in higher organisms, in bacteria, they are known to include enzymes termed transposases, as well as molecules which regulate expression of the transposase and other genes encoding transposition-specific proteins.

Transposable elements are mobile stretches of DNA which are defined by two end terminals, usually denoted attL and attR at the left and right attachment ends respectively. Natural transposable elements contain DNA coding for transposases, and often portable genes conferring traits such as resistance to antibiotics. Some transposable elements can insert randomly into targeted DNA while others are sequence specific in their insertion sites. Transposons have been implicated as having a major role in evolution, and there is evidence for natural multifunctional enzymes having originated from the natural fusion of different protein domains.

Scientists have taken advantage of transposons to transport reporter genes for use in studying gene expression. These include transcriptional (Type I) fusions and translational (Type II) fusions. Transcriptional fusions, unlike translational fusions, place a reporter gene under the control of another promoter, but do not translationally fuse two protein domains. Translational fusions have generally been made to link a reporter gene carried inside the transposon to the translational frame of the target gene so that the reporter gene is expressed under direct control of the transcription and translation signals of the target gene of interest to study gene regulation. This requires that an open reading frame extend through the end of the transposable element to join an internal reporter protein to external translational sequences. This usually results in complete inactivation of the target gene.

4.2.2. Rational protein design

The main goal of modern protein chemistry is to be able to design proteins with desired functions. The approach taken by the majority of scientists has been called rational protein design, which is highly dependent on knowledge of protein structure in three dimensions and protein folding. While an extremely powerful tool, rational protein design is currently limited to a small subset of enzymes: those with well defined three dimensional structures. The classic examples include proteins such as subtilisin from several *Bacillus* sp. (Wells, Powers et al. Proc Natl Acad Sci USA. (1987). 84: 1219-1223.) and T4 lysozyme

(Matsumura, Becktel et al. (1989). *Proc Natl Acad Sci USA*. 86: 6562-6566.). The three dimensional organization of the amino acid side chains of the protein must be known at high resolution, and often protein/substrate and mutant structures must be known as well. While this method is useful, it is expensive, time consuming and requires difficult predictions. Even with a complete three dimensional X-ray crystallography structure in hand, it has proven difficult to design proteins with specific desired activities. To address these problems, methods which allow the accelerated evolution of proteins in vivo or in vitro can take advantage of natural mutagenic mechanisms and their resulting variability.

4.2.2.1. Protein fusions

Protein (translational) fusion can be created by the joining of translational sequences from two different genes to create a hybrid protein molecule. These applications have traditionally included the study of gene expression in microorganisms and eukaryotes (Casadaban, Martinez-Arias et al. (1983). *Recombinant DNA. Methods in Enzymology*. 100: 293-308.).

The use of protein fusion in vitro to generate hybrid (chimeric) proteins has gained importance in the development of novel or multifunctional enzyme activities. Applications include using protein domains to aid in protein purification (Sherwood. (1991). *TIBTECH*. 9: 1- 3.) or to tag proteins for delivery to specific cellular locations (Crozel, Lazdunski et al. (1984). *FEBS Lett*. 172: 183-188; Moore and Kelly. (1986). *Nature*. 321: 443-446; Roitsch and Lehle. (1991). *Eur J Biochem*. 195: 145-50.). Domain shuffling between different proteins shows promise to create protein products with unique uses, often to bind a particular enzymatic activity to a site of interest (Panayotatos, Fontaine et al. (1988). *Molecular genetics of bacteria and phages: prokaryotic gene regulation*. 174.), as a reporter system for protein- protein interactions (Fields and Song. (1989). *Nature*. 340: 245-246.), to study the relatedness of different proteins (Caramori, Albertini et al. (1991). *Gene*. 98: 37-44.) or as targeted pharmaceuticals (Pastan and FitzGerald. (1991). *Science*. 254: 1173-1177.). Finally, another promising application of protein fusion to biotechnology is in creating multi-catalytic enzymes which are important in biocatalysis since they represent an alternative to co-immobilization and chemical crosslinking to create multienzyme systems (Bulow and Mosbach. (1991). *TIBTECH*. 9: 226-231.).

4.2.2.2. Problems in the development of protein fusions

Unfortunately, the construction of functional hybrid proteins can require an extensive knowledge of a protein's structure and functional domains in order to select a proper site for fusion. Many attempts have failed to produce the desired properties (Bowie and Sauer. (1989). *J Bio Chem.* 264: 7596- 7602; Ellis, Morgan et al. (1986). *Proc Natl Acad Sci, USA.* 83: 8137- 8141; Guan and Rose. (1984). *Cell.* 37: 779-787; Hellebust, Murby et al. (1989). *BioTech.* 7: 165-168.). Random deletions can be made to fuse two domains, but this is typically done for only one domain at a time, and the cost and time involved in such trial and error efforts can be substantial. In addition, while some gene fusions can be used to stabilize proteins, unstable structures are often formed which are recognized by the cellular degrading machinery (Bowie and Sauer. (1989). *J Bio Chem.* 264: 7596-7602; Hellebust, Murby et al. (1989). *BioTech.* 7: 165-168.). Also, even with the advanced level of molecular biological techniques available today, cloning remains a labor-intensive procedure, the results of which are not trivially predictable.

Several tools have been developed to make the construction of protein fusions simpler. These tools include new plasmid systems with convenient restriction sites (Shapira, Chou et al. (1983). *Gene.* 25: 71- 82.), and a method for making gene fusions using the Polymerase Chain Reaction (PCR) so that convenient restriction sites are not required (Horton, Hunt et al. (1989). *Gene.* 77: 61-8.). None of these approaches, however, offers a truly simple way of making random protein fusions which eliminates the labor-intensive, trial and error aspects of traditional techniques, especially in the case when at least one of the two domains being studied has not yet been well characterized.

4.2.3. In Vitro Transposition Systems

In vitro transposition systems that utilize the particular transposable elements of bacteriophage Mu and bacterial transposon Tn10 have been described, by the research groups of Kiyoshi Mizuuchi and Nancy Kleckner, respectively.

The bacteriophage Mu system was first described by Mizuuchi, K., "In Vitro Transposition of Bacteria Phage Mu: A Biochemical Approach to a Novel Replication

Reaction," *Cell*:785-794 (1983) and Craigie, R. et al., "A Defined System for the DNA Strand-Transfer Reaction at the Initiation of Bacteriophage Mu Transposition: Protein and DNA Substrate Requirements," *P.N.A.S. U.S.A.* 82:7570-7574 (1985). The DNA donor substrate (mini-Mu) for Mu in vitro reaction normally requires six Mu transposase binding sites (three of about 30 bp at each end) and an enhancer sequence located about 1 kb from the left end. The donor plasmid must be supercoiled. Proteins required are Mu-encoded A and B proteins and host-encoded HU and IHF proteins. Lavoie, B. D, and G. Chaconas, "Transposition of phage Mu DNA," *Curr. Topics Microbiol. Immunol.* 204:83-99 (1995). The Mu-based system is disfavored for in vitro transposition system applications because the Mu termini are complex and sophisticated and because transposition requires additional proteins above and beyond the transposase.

The Tn10 system was described by Morisato, D. and N. Kleckner, "Tn10 Transposition and Circle Formation in vitro," *Cell* 51:101-111 (1987) and by Benjamin, H. W. and N. Kleckner, "Excision Of Tn10 from the Donor Site During Transposition Occurs By Flush Double-Strand Cleavages at the Transposon Termini," *P.N.A.S. U.S.A.* 89:4648-4652 (1992). The Tn10 system involves a supercoiled circular DNA molecule carrying the transposable element (or a linear DNA molecule plus *E. coli* IHF protein). The transposable element is defined by complex 42 bp terminal sequences with IHF binding site adjacent to the inverted repeat. In fact, even longer (81 bp) ends of Tn10 were used in reported experiments. Sakai, J. et al., "Identification and Characterization of Pre-Cleavage Synaptic Complex that is an Early Intermediate in Tn10 transposition," *E.M.B.O. J.* 14:4374-4383 (1995). In the Tn10 system, chemical treatment of the transposase protein is essential to support active transposition. In addition, the termini of the Tn10 element limit its utility in a generalized in vitro transposition system.

Both the Mu- and Tn10-based in vitro transposition systems are further limited in that they are active only on covalently closed circular, supercoiled DNA targets. What is desired is a more broadly applicable in vitro transposition system that utilizes shorter, more well defined termini and which is active on target DNA of any structure (linear, relaxed circular, and supercoiled circular DNA).

4.2.4. Importance Of Transposons In Agriculture

Currently, there is a great deal of interest in the development of gene transfer vectors for use with agriculturally important plants (See Outlook for Science and Technology, The Next Five Years, Vol. III (National Science Foundation (1982); and O.T.A. Report, Impact of Applied Genetics (1981)).

Although the United States presently has an excess productivity in the agricultural sector, this is recognized as a local and short term condition. Thus, agricultural research and planning must be based on long term considerations. The variety of problems surrounding increases in population, degradation of prime farm land and decreasing availability of good farm land necessitates the increased use of marginal land, as well as exogenous fertilizers and chemical pest control compositions.

Classical plant breeding programs have thus far been successful in increasing agricultural productivity. However, a substantial fraction of the increase in farm productivity experienced in the United States in the past 40 years is attributable to the use of fertilizers and modern energy intensive cultivation practices, both of which are increasingly costly. The ability of plant breeding alone to sustain productivity is a matter of some question. Plant breeders are divided in their views on whether genetic improvements will continue at the rate that has occurred over the past few decades or will begin to level out. Since such questions cannot be resolved a priori, it is prudent to explore a variety of additional means by which agronomically useful traits can be accumulated and improved in major crop plants. The unconventional areas that are presently receiving the most attention in the academic research establishment, as well as in both small and large firms with plant-oriented research programs are wide genetic crosses, tissue culture and the development of gene transfer systems that circumvent fertility barriers.

In the past, many attempts have been made to transform plant cells with DNA from a variety of sources. The first unequivocal demonstration that DNA transfer can and does occur in plants emerged from the work described above on *Agrobacterium tumefaciens* Ti plasmid. However, Ti-plasmid mediated gene transfer is presently accomplished only in dicotyledonous plants that interact with the plasmid's natural host bacterium. Since most major crop species are monocotyledonous, ti-plasmid mediated gene transfer has limited applications.

4.2.4.1. Use Of Transposons On The Ti Plasmid Of *Agrobacterium*

In higher organisms, transposons have been, or are being, used in several ways. For example, transposons are used as mutagens on the Ti plasmid of *Agrobacterium tumefaciens*. That is, a method for using bacterial transposons to cause insertion mutations in the *Agrobacterium tumefaciens* Ti plasmid, the causative agent of crown gall disease in dicotyledonous plants, has been developed. (See Zambriski, P., Goodman, H., Van Montagu, M. and Schell, J., *Mobile Genetic Elements*, J. Shapiro, Ed., (Academic Press) New York, pp. 506-535 (1983)). By this technique, it has become possible to identify the plasmid-borne genes that are responsible for virulence, as well as those that are responsible for the tumorous transformation of plant cells caused by the Ti plasmid. Further, it has become possible to show by using transposable elements, that a portion of the Ti plasmid can be integrated into plant genomes and can act as a vehicle for transferring genes from virtually any organism to any dicotyledonous plant that is susceptible to *Agrobacterium tumefaciens*.

4.2.4.2. Use Of Transposons In Maize

In maize, a monocotyledon, transposable elements were first genetically identified in the mid-1940s. These elements have been studied extensively and their genetic behavior has been extensively reviewed (See McClintock, B., *Cold Spring Harbor Symp. Quant. Biol.* 16:13-47 (1951); McClintock, B., *Cold Spring Harbor Symp. Quant. Biol.* 21:197-216 (1956); McClintock, B., *Brookhaven Symp. Biol.* 18:162-184 (1965); Fincham, J. R. S., and Sastry, G. R. K., *Ann. Rev. Genet.* 8:15-50 (1974); and Fedoroff, N., *Mobile Genetic Elements*, J. Shapiro, Ed., (Academic Press) New York, pp. 1-63 (1983)).

It has been demonstrated that transposons are normal, although cryptic, residents of the maize genome and that upon activation, they are responsible for various types of genetic rearrangements, including chromosome breakage, deletions, duplications, inversions and translocations. In addition, it has been shown that certain common types of unstable mutations, which have been studied for decades in both maize and in other organisms, are attributable to the insertion of transposons into genes or genetic loci.

4.2.5. Mu and the Transposing Bacteriophage

Bacteriophage Mu represents a class of transposons known as transposable bacteriophage which both function as a virus and a transposon. Mu replicates itself by transposing at high frequency, but can also integrate randomly into its host's genome as a lysogen. Mu is a model system for other transposable bacteriophage which are generally highly homologous. These include the *Pseudomonas* phage D3112, D108, and several other phage.

Because of the randomness of Mu insertions, and the high levels of transposition which can be generated by Mu strains containing a temperature sensitive transposition repressor (Mu_{ts} strains), Mu has been developed into a genetic tool to study gene expression in bacterial systems. Transposition of Mu derivatives has allowed scientists to perturb and examine the basic components critical to protein expression and translation. The most commonly used Mu derivatives include reporter genes which have been integrated into the Mu genome.

4.2.5.1. Type I Fusions

Type I transcriptional fusions have been used to study gene expression and regulation by co-opting the native transcriptional signal to express the exogenous reporter gene. For example for gene expression in *E. coli*, yeast, and *Drosophila* development.

4.2.5.2. Type II Fusions

Type II fusions have also been used to study gene expression and regulation, but in this case not only co-opt the transcriptional signals, but any translational signals as well to express the reporter gene. In this type of system the protein product usually only expresses the activity of the reporter exogenous gene.

MudII elements are mini-Mu deletion elements which are type II Mu transposable elements. Examples of these include beta-galactosidase fusion elements, where a beta-galactosidase (*lacZ*) reporter gene is inserted via transposable elements to detect transcription and translation of regulated gene systems. This usually results in the inactivation of the targeted gene.

Two types of Mu protein fusions have been developed, *lacZ* fusion elements and *nptI* fusion elements (Symonds, Toussaint et al. (1987). Phage Mu) The *lacZ* elements have been used to study translation regulation, determination of the translation phase of target genes, infer the location of a protein fusion by hybrid protein size, determine amino terminal sequence, and raise antibodies to regions of the protein of interest. By far the major goal of these studies has been to determine mechanisms of gene expression in the studied organisms.

The *nptI* system was designed to perform transposon-tagging since *nptI* is known to function as an aminoglycoside resistance gene in a variety of organisms. Transposon tagging is a method of creating a mutant by inserting a transposon with a selectable marker into the gene of interest so that mutants which inactivate the gene can be identified and maintained. This element is useful since it allows the *nptI* to be directly linked to the transcription/translation system of the organism being studied.

In these studies there has been no emphasis on creating novel proteins with new activities using these transposable elements. More importantly, these Mu elements are restricted to making amino-terminal fusions to the reporter protein. In these cases the inserted reporter gene is fused to the carboxy-end of the truncated targeted protein, terminating inside the Mu. If the transposable element were to insert before the amino terminal of a targeted gene, functional translation could only occur on the marker gene by itself, and no translation of the target gene would occur.

4.2.5.3. Problems with Mu

Unfortunately, available Mu elements had several problems. First, it has not been demonstrated that Mu elements can be readily used as a general method for the development of fusion proteins with two active domains. Second, the Mu elements used thus far for creation of protein fusions can not be used for construction of "carboxy-terminal" fusions since they did not have an open reading frame extending into the element. Third, the Mu elements previously used have long linker regions which incorporate a 40 amino acid linker between the fused domains. This could create protein folding problems or unwanted domain interactions. Fourth the currently existing Mu elements had only a single restriction site for the insertion of protein domains. Finally, although Mu elements which had deleted ends existed, it was not known whether they would transpose well with additional sequences added in such close proximity to the right end and whether the intervening linker region which would join the two protein domains would interfere with the construction of active chimeric proteins.

4.2.6. Other transposons

Other transposons have been used in a similar manner as Mu to create lac fusions to study gene expression. These include Tn10 and Tn917 (Berg and Howe. (1989). Mobile DNA). The Tn5 element has also been used to construct phoA fusions in vivo. Fusions with alkaline phosphatase (phoA) have also been used to probe the structure of membrane bound proteins (Lloyd and Kadner. (1990). J Bacteriol. 172: 1688-93.). In general, these transposons have been used to study the membrane topology structure of a particular gene and protein secretion. The resultant fusion proteins are also limited to amino-terminal

fusion of the reporter PhoA reporter protein resulting in fusion at the carboxy end of the targeted gene.

In general, these types of fusions have been applied to the study of gene expression. These elements were constructed with truncated marker proteins that extend through the end of the transposon. Transposition of the element can create an in-frame fusion with a target gene, thereby activating expression. Mini-Mu elements are used because they transpose at high frequencies, insert randomly, and can be packaged along with a target plasmid and transduced to a new cell (Symonds, Toussaint et al. (1987). Phage Mu). Some of the more pertinent work that has been done in the area of transposable elements are detailed in the following.

Namgoong et al., (1994), teach that the Mu transposition reaction attachment sites attL and attR can promote the assembly of higher order complexes held together by non-covalent protein-DNA and protein-protein interactions. (Namgoong, Jayaram et al. (1994). J Mol Biol. 238: 514-527.)

Harel et al., (1990), teach that in Mu helper-mediated transposition packaging the left end contains an essential domain defined by nucleotides 1 to 54 of the left end (attL). At the right end (attR), they teach that the essential sequences for transposition require not more than the first 62 base pairs (bp), although the presence of sequences between 63 and 117 bp from the right end increase transposition frequency about 15-fold. (Harel, Dupliessis et al. (1990). Arch Microbiol. 154: 67-72.)

Groenen and van de Putte (1986), teach that the Mu A protein binds weakly to sequences between nucleotides 1 to 30 on the right end (R1) and between nucleotides 110 and 135 on the left end (L2). Mutations in these weak A binding sites have a greater effect on transposition than mutations of corresponding base pairs in the stronger A binding sites, located adjacent to these weak A binding sites. (Groenen and van de Putte. (1986). J Mol Biol. 189: 597-602.)

Groenen and et al. (1985) teach the DNA sequences at the end of the genome of bacteriophage Mu that are essential for transposition. (Groenen, Timmers et al. (1985). Proc Natl Acad Sci, USA. 82: 2087-2091.)

Lloyd and Kadner teach the how to probe the topology of the uhpT sugar phosphate transporter using a Tn5phoA element. (Lloyd and Kadner. (1990). J Bacteriol. 172: 1688-93.)

Phage Mu (1987), Cold Spring Harbor Laboratory Press (Symonds, et al eds.) teaches general methods for handling and working with bacteriophage Mu as a transposon, and describes the various uses of mini- Mu elements including the construction of Mu transcriptional and translational fusions.

Silhavy and Beckwith (1985) teaches the various uses of lac fusions for the study of biological problems. (Silhavy and Beckwith. (1985). Microbiol Rev. 49: 398-418.)
Mobile DNA, (1989), American Society for Microbiology, Publishers. (Berg, Howe, eds) describes transposons.

Casadaban, et al. (1983) Methods in Enzymol, provides a good general review of beta-galactosidase gene fusions for the study of gene expression. (Casadaban, Martinez-Arias et al. (1983). Recombinant DNA. Methods in Enzymology. 100: 293-308.)

4.3.1. In Vitro Transposition System

The present invention is summarized in that an in vitro transposition system comprises a preparation of a suitably modified transposase of bacterial transposon Tn5, a donor DNA molecule that includes a transposable element, a target DNA molecule into which the transposable element can transpose, all provided in a suitable reaction buffer.

4.3.1.1. Donor DNA Molecule: Transposable DNA Sequence Of Interest

The transposable element of the donor DNA molecule is characterized as a transposable DNA sequence of interest, the DNA sequence of interest being flanked at its 5'-and 3'-ends by short repeat sequences that are acted upon in trans by Tn5 transposase.

4.3.1.1.1. Modified Transposase Enzyme Comprises Two Classes Of Differences From Wild Type Tn5 Transposase

The invention is further summarized in that the suitably modified transposase enzyme comprises two classes of differences from wild type Tn5 transposase, where each class has a separate measurable effect upon the overall transposition activity of the enzyme and where a greater effect is observed when both modifications are present. The suitably modified enzyme both (1) binds to the repeat sequences of the donor DNA with greater avidity than wild type Tn5 transposase ("class (1) mutation") and (2) is less likely than the wild type protein to assume an inactive multimeric form ("class (2) mutation"). A suitably modified Tn5 transposase of the present invention that contains both class (1) and class (2) modifications induces at least about 100-fold ($\pm 10\%$) more transposition than the wild type enzyme, when tested in combination in an in vivo conjugation assay as described by Weinreich, M. D., "Evidence that the cis Preference of the Tn5 Transposase is Caused by Nonproductive Multimerization," *Genes and Development* 8:2363-2374 (1994), incorporated herein by reference. Under optimal conditions, transposition using the modified transposase may be higher. A modified transposase containing only a class (1) mutation binds to the repeat sequences with sufficiently greater avidity than the wild type Tn5 transposase that such a Tn5 transposase induces about 5- to 50-fold more transposition than the wild type enzyme, when measured in vivo. A modified transposase containing only a class (2) mutation is sufficiently less likely than the wild type Tn5

transposase to assume the multimeric form that such a Tn5 transposase also induces about 5- to 50-fold more transposition than the wild type enzyme, when measured *in vivo*.

4.3.1.1.2. Method For Transposing The Transposable Element

In another aspect, the invention is summarized in that a method for transposing the transposable element from the donor DNA into the target DNA *in vitro* includes the steps of mixing together the suitably modified Tn5 transposase protein, the donor DNA, and the target DNA in a suitable reaction buffer, allowing the enzyme to bind to the flanking repeat sequences of the donor DNA at a temperature greater than 0degree C, but no higher than about 28 °C, and then raising the temperature to physiological temperature (about 37 °C) whereupon cleavage and strand transfer can occur.

It is an object of the present invention to provide a useful *in vitro* transposition system having few structural requirements and high efficiency.

It is another object of the present invention to provide a method that can be broadly applied in various ways, such as to create absolute defective mutants, to provide selective markers to target DNA, to provide portable regions of homology to a target DNA, to facilitate insertion of specialized DNA sequences into target DNA, to provide primer binding sites or tags for DNA sequencing, to facilitate production of genetic fusions for gene expression studies and protein domain mapping, as well as to bring together other desired combinations of DNA sequences (combinatorial genetics).

4.3.1.1.3. Modified Transposase Advantages Over Wild Type Tn5 Transposase

It is a feature of the present invention that the modified transposase enzyme binds more tightly to DNA than does wild type Tn5 transposase.

It is an advantage of the present invention that the modified transposase facilitates *in vitro* transposition reaction rates of at least about 100-fold higher than can be achieved using wild type transposase (as measured *in vivo*). It is noted that the wild-type Tn5 transposase shows no detectable *in vitro* activity in the system of the present invention. Thus, while it is difficult to calculate an upper limit to the increase in activity, it is clear that hundreds, if not thousands, of colonies are observed when the products of *in vitro* transposition are assayed *in vivo*.

It is another advantage of the present invention that in vitro transposition using this system can utilize donor DNA and target DNA that is circular or linear.

It is yet another advantage of the present invention that in vitro transposition using this system requires no outside high energy source and no other protein other than the modified transposase.

Other objects, features, and advantages of the present invention will become apparent upon consideration of the following detailed description.

4.3.2. Transposable elements for generating functional fusion proteins

The instant invention encompasses transposable elements for generating functional fusion proteins comprising at least a left attachment site and a right attachment site where there is an open reading frame in both directions through at least one of the ends. In one embodiment the instant invention encompasses a transposable element containing a polylinker that is located adjacent to one end of the transposable element, while still allowing for open translational reading frames to enter and exit the near end. In another embodiment, the invention encompasses a transposable element as in containing a polylinker with an inserted exogenous DNA sequence. This exogenous DNA can encode for a complete protein, a functional portion of a protein, an expressible or inducible segment of DNA, or any other suitable DNA sequence. Also encompassed by the instant invention is a transposable element, where the expression of the exogenous DNA is under controlled regulation of a promoter, enhancer, or repressor which target sequences can be a part of the exogenous DNA segment, or a part of the transposable element construct. In a preferred embodiment, the instant invention encompasses a transposable element of containing a right attachment site of 50 to 62 nucleotides. The instant invention encompasses a transposable element for generating functional fusion proteins comprising at least a left attachment site and a right attachment site where there is an open translational reading frame extending out through one of the ends of the element. The instant invention also encompasses, in one embodiment, full-length attachment sites, in which translational open reading frame interrupting stop codons, which prevent the reading frame from extending out of the transposable element through a attachment site that is within 400 bases, have been removed by selective substitution of nucleotides in the

nucleic acid sequence. Thus the instant invention also encompasses a transposable element for generating functional fusion proteins comprising at least a left attachment site and a right attachment site where the endogenous stop codons that would prevent translation through the ends have been removed by point mutation.

4.3.2.1. Transposable element containing the Protein A domain

In another particular embodiment of constructs of the instant invention is a transposable element containing the Protein A domain which allows functional protein fusion with a target protein. The instant invention also embodies a transposable element which allows for functional carboxy or amino terminal fusion of the Protein A domain to a targeted protein sequence. Thus the instant invention provides for methods of making Protein A fusion proteins which are either amino- or carboxy-terminal fusion proteins.

4.3.2.2. A cell wherein a transposable element for generating functional fusion proteins

The methods and constructs of the instant invention also provide for a cell wherein a transposable element for generating functional fusion proteins, comprising at least a left attachment site and a right attachment site with a protein domain whose translational open reading frame extends through one of the ends of the element, is integrated into the cell genome. In one embodiment, the transposable element is integrated into an autonomously replicating DNA form within a cell of the instant invention.

The instant invention encompasses a method for generating functional fusion proteins, with a transposable element which contains at least a left attachment site and a right attachment site where there is an open reading frame in both directions through at least one of the ends, comprising insertion of an exogenous DNA sequence into the transposable element, transposing into a target DNA sequence with the transposable element, detecting the presence of the exogenous DNA insert, and selecting for the presence of the exogenous DNA insert. Another embodiment of the instant invention encompasses a method for in vitro or in vivo protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site, a polylinker,

and a right attachment site of between 50 and 62 nucleic acid base pairs, inserting within the polylinker an exogenous DNA sequence, transposing the transposable element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA. In particular the instant invention encompasses a right attachment site of 58 nucleotides in length.

4.3.2.3. In vitro or in vivo carboxy terminal translational protein fusion in a target organism genome

Another particular embodiment of the instant invention encompasses a method for in vitro or in vivo carboxy terminal translational protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site and a right attachment site of between 50 and 62 nucleic acid base pairs, inserting within the transposable element an exogenous DNA sequence, transposing the transposable element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA. In a particular embodiment the right attachment site is 58 nucleotides in length. The instant invention also provides for a method for in vitro or in vivo amino terminal translational protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site attL and a right attachment site attR of between 50 and 62 nucleic acid base pairs, inserting within the transposable element an exogenous DNA sequence, transposing the transposable element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA. In a preferred embodiment of the instant invention, the attachment sites are derived from the Mu family of transposable elements. In another preferred embodiment of the instant invention, the right attachment site is 58 nucleotides in length.

4.3.2.4. A plasmid for inserting target sequences for protein fusion

The instant invention also encompasses a plasmid for inserting target sequences for protein fusion which comprises in the following order, a transcription termination site, an extended translational target region with no stop codons, and a polylinker sequence for cloning protein domains into containing no stop codons in frame with the translational

target region. In another embodiment, the plasmid additionally contains a screenable marker fused to the translational target region, and the polylinker reading frame located after the polylinker region.

Thus the instant invention teaches the use of transposable elements with deletions in an end terminal attachment site which result in multiple reading frames in both directions through the end, and allows for the generation of functional protein fusions. In one particular embodiment the particular end is the right end attachment site of Mu known as attR. The invention specifically teaches constructs which allow for carboxy-terminal and/or amino-terminal fusion events to occur. In one embodiment, the instant invention encompasses a transposable element comprising a left attachment site attL, and a right attachment site of no more than 58 nucleic acid base pairs which will transpose an exogenous DNA sequence into a target DNA sequence. In a preferred embodiment, the attachment sites are derived from the Mu family of transposable elements. The teachings of the instant invention can also be applied to create similar constructs which allow for similar translation reading frames via the left attachment site. The instant invention also encompasses transposable elements comprising a left attachment site attL, a polylinker, and a right attachment site. In this embodiment, the polylinker allows for multiple restriction sites which allow for the convenient insertion of exogenous DNA segments. In a preferred embodiment, the transposable element is constructed such that it allows for open reading frames in both orientations. In a further embodiment, the exogenous DNA inserted within the transposable element is under controlled expression by a promoter, enhancer or repressor element. In a preferred embodiment, the instant invention encompasses a transposable element comprising a left attachment site attL, a polylinker region, and a right attachment site of no more than 58 nucleic acid base pairs derived from the Mu family of transposable elements, which will transpose an exogenous DNA sequence into a target DNA sequence and allow for open reading frames in both orientations.

4.3.2.5. Transposable element comprising a left attachment site

The instant invention provides for the use of a transposable element comprising a left attachment site attL, an exogenous DNA, and a right attachment site for generating in vivo

protein fusions in a target DNA. The target DNA can be plasmid DNA, DNA segments, genomic DNA, or other DNA targets. In a preferred embodiment of the instant invention, the right end of the transposable element consists of no more than 58 nucleic acid base pairs derived from the Mu family of transposable elements. The instant invention encompasses methods for using transposable element constructs with deletions in the right end for generating fusion proteins in vitro and in vivo. A method of the instant invention allows for the rapid and efficient generation of alternatively fused proteins suitable for screening for activity. In one embodiment of the method of the instant invention generates fusion proteins in which the exogenous DNA segment has been transposed such that the resultant fusion protein functionally expresses the exogenous protein as a amino- terminal fusion to the carboxy end of a targeted protein. In a further embodiment of a method of the instant invention, the transposable element causes the insertion of the exogenous DNA such that the functionally expressed fusion protein consists of the exogenous protein at the amino end, linked to the endogenous protein at the carboxy end of the fusion protein.

The instant invention provides for a method for in vivo protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site attL, a polylinker, and a right attachment site of no more than 58 nucleic acid base pairs, inserting within the polylinker an exogenous DNA sequence, transposing the transposable element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA. In a preferred embodiment, the instant invention provides for the in vivo carboxy terminal translational protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site attL and a right attachment site attR of no more than 58 nucleic acid base pairs, inserting within the transposable element an exogenous DNA sequence, transposing the transposable element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA. In another preferred embodiment, the instant invention provides for in vivo amino terminal translational protein fusion in a target organism genome comprising constructing a transposable element with a left attachment site attL and a right attachment site attR of no more than 58 nucleic acid base pairs, inserting within the transposable element an exogenous DNA sequence, transposing the transposable

element into a target organism genome, expanding the target organism, isolating protein, and screening for fusion protein containing the translated exogenous DNA.

4.3.2.6. Transposable elements in which the stop codons have been removed

In a further embodiment of the instant invention, the method for generating novel fusion proteins encompasses the use of transposable elements in which the stop codons present in the attR region have been removed by point mutation such that there is an open reading frame leading out of the transposable element, allowing for the amino-terminal or carboxy-terminal fusion of the exogenous protein with the targeted protein by a linker which is translated from the attR region.

The constructs and methods of the instant invention are useful for the rapid and efficient generation of functional fusion proteins. The constructs and methods of the instant invention are useful in that they reduce the labor intensive burdens that accompany generation of protein fusions by traditional molecular cloning techniques.

The following descriptions and examples are meant only by way of illustration of the instant invention, and are in no way intended to limit the scope of the instant invention. One with ordinary skill in the art will be able to understand and use the descriptions of the instant specification to use all of the embodiments which are contemplated and encompassed by the constructs and methods of the instant invention.

4.4. Transposons - Specialized Applications

4.4.1. In Vitro System For Introducing Any Transposable Element From A Donor DNA Into A Target DNA

It will be appreciated that this technique provides a simple, in vitro system for introducing any transposable element from a donor DNA into a target DNA. It is generally accepted and understood that Tn5 transposition requires only a pair of OE termini, located to either side of the transposable element. These OE termini are generally thought to be 18 or 19 bases in length and are inverted repeats relative to one another. Johnson, R. C., and W. S. Reznikoff, Nature 304:280 (1983), incorporated herein by reference. The Tn5 inverted repeat sequences, which are referred to as "termini" even though they need not be at the termini of the donor DNA molecule, are well known and understood.

Apart from the need to flank the desired transposable element with standard Tn5 outside end ("OE") termini, few other requirements on either the donor DNA or the target DNA are envisioned. It is thought that Tn5 has few, if any, preferences for insertion sites, so it is possible to use the system to introduce desired sequences at random into target DNA. Therefore, it is believed that this method, employing the modified transposase described herein and a simple donor DNA, is broadly applicable to introduce changes into any target DNA, without regard to its nucleotide sequence. It will, thus, be applied to many problems of interest to those skilled in the art of molecular biology.

4.4.1.1. Modified Transposase Protein Is Combined In A Suitable Reaction Buffer With The Donor DNA And The Target DNA

In the method, the modified transposase protein is combined in a suitable reaction buffer with the donor DNA and the target DNA. A suitable reaction buffer permits the transposition reaction to occur. A preferred, but not necessarily optimized, buffer contains spermidine to condense the DNA, glutamate, and magnesium, as well as a detergent, which is preferably 3-[(3-cholamidopropyl) dimethyl-ammonio]-1-propane sulfonate ("CHAPS"). The mixture can be incubated at a temperature greater than 0 °C and as high as about 28 °C to facilitate binding of the enzyme to the OE termini. A preferred temperature range is between 16° C and 28 ° C A most preferred pretreatment temperature is about 20 °C Under different buffer conditions, however, it may be possible to use other

below-physiological temperatures for the binding step. After a short pretreatment period of time (which has not been optimized, but which may be as little as 30 minutes or as much as 2 hours, and is typically 1 hour), the reaction mixture is diluted with 2 volumes of a suitable reaction buffer and shifted to physiological conditions for several more hours (say 2-3 hours) to permit cleavage and strand transfer to occur. A temperature of 37 °C, or thereabouts, is adequate. After about 3 hours, the rate of transposition decreases markedly. The reaction can be stopped by phenol-chloroform extraction and can then be desalted by ethanol precipitation.

Following the reaction and subsequent extraction steps, transposition can be assayed by introducing the nucleic acid reaction products into suitable bacterial host cells (e.g., *E. coli* K-12 DH5 α cells (recA-); commercially available from Life Technologies (Gibco-BRL)) preferably by electroporation, described by Dower et al., Nuc. Acids. Res. 16:6127 (1988), and monitoring for evidence of transposition, as is described elsewhere herein.

Those persons skilled in the art will appreciate that apart from the changes noted herein, the transposition reaction can proceed under much the same conditions as would be found in an in vivo reaction. Yet, the modified transposase described herein so increases the level of transposition activity that it is now possible to carry out this reaction in vitro where this has not previously been possible. The rates of reaction are even greater when the modified transposase is coupled with an optimized buffer and temperature conditions noted herein.

4.4.1.1.1. Preparation Of A Modified Tn5 Transposase Enzyme That Differs From Wild Type Tn5 Transposase

In another aspect, the present invention is a preparation of a modified Tn5 transposase enzyme that differs from wild type Tn5 transposase in that it (1) binds to the repeat sequences of the donor DNA with greater avidity than wild type Tn5 transposase and (2) is less likely than the wild type protein to assume an inactive multimeric form. An enzyme having these requirements can be obtained from a bacterial host cell containing an expressible gene for the modified enzyme that is under the control of a promoter active in the host cell. Genetic material that encodes the modified Tn5 transposase can be introduced (e.g., by electroporation) into suitable bacterial host cells capable of supporting expression of the genetic material. Known methods for overproducing and preparing other

Tn5 transposase mutants are suitably employed. For example, Weinreich, M. D., et al., supra, describes a suitable method for overproducing a Tn5 transposase. A second method for purifying Tn5 transposase was described in de la Cruz, N. B., et al., "Characterization of the Tn5 Transposase and Inhibitor Proteins: A Model for the Inhibition of Transposition," J. Bact. 175:6932-6938 (1993), also incorporated herein by reference. It is noted that induction can be carried out at temperatures below 37 °C, which is the temperature used by de la Cruz, et al. Temperatures at least in the range of 33 to 37 °C are suitable. The inventors have determined that the method for preparing the modified transposase of the present invention is not critical to success of the method, as various preparation strategies have been used with equal success.

Alternatively, the protein can be chemically synthesized, in a manner known to the art, using the amino acid sequence attached hereto as SEQ ID NO:2 as a guide. It is also possible to prepare a genetic construct that encodes the modified protein (and associated transcription and translation signals) by using standard recombinant DNA methods familiar to molecular biologists. The genetic material useful for preparing such constructs can be obtained from existing Tn5 constructs, or can be prepared using known methods for introducing mutations into genetic material (e.g., random mutagenesis PCR or site-directed mutagenesis) or some combination of both methods. The genetic sequence that encodes the protein shown in SEQ ID NO:2 is set forth in SEQ ID NO:1.

The nucleic acid and amino acid sequence of wild type Tn5 transposase are known and published. N.C.B.I. Accession Number U00004 L19385, incorporated herein by reference.

4.4.1.1.2. Differences Observed In Modified Tn5 Transposase Enzyme From Wildtype

4.4.1.1.2.1. Higher Binding Preference Of The Transposase For Outside End ("OE") Termini

In a preferred embodiment, the improved avidity of the modified transposase for the repeat sequences for OE termini (class (1) mutation) can be achieved by providing a lysine residue at amino acid 54, which is glutamic acid in wild type Tn5 transposase. The mutation strongly alters the preference of the transposase for OE termini, as opposed to inside end ("IE") termini. The higher binding of this mutation, known as EK54, to OE

termini results in a transposition rate that is about 10-fold higher than is seen with wild type transposase. A similar change at position 54 to valine (mutant EV54) also results in somewhat increased preference (about 3-fold higher than wild type) for OE termini, as does a threonine-to-proline change at position 47 (mutant TP47; about 10-fold higher). It is believed that other, comparable transposase mutations (in one or more amino acids) that increase binding avidity for OE termini may also be obtained which would function as well or better in the in vitro assay described herein.

One of ordinary skill will also appreciate that changes to the nucleotide sequences of the short repeat sequences of the donor DNA may coordinate with other mutation(s) in or near the binding region of the transposase enzyme to achieve the same increased binding effect, and the resulting 5- to 50-fold increase in transposition rate. Thus, while the applicants have exemplified one case of a mutation that improves binding of the exemplified transposase, it will be understood that other mutations in the transposase, or in the short repeat sequences, or in both, will also yield transposases that fall within the scope and spirit of the present invention. A suitable method for determining the relative avidity for Tn5 OE termini has been published by Jilk, R. A., et al., "The Organization of the Outside end of Transposon Tn5," J. Bact. 178:1671-79 (1996).

4.4.1.1.2.2. Less Likely Than The Wild Type Protein To Assume An Inactive Multimeric Form

The transposase of the present invention is also less likely than the wild type protein to assume an inactive multimeric form. In the preferred embodiment, that class (2) mutation from wild type can be achieved by modifying amino acid 372 (leucine) of wild type Tn5 transposase to a proline (and, likewise by modifying the corresponding DNA to encode proline). This mutation, referred to as LP372, has previously been characterized as a mutation in the dimerization region of the transposase. Weinreich, et al., *supra*. It was noted by Weinreich et al. that this mutation at position 372 maps to a region shown previously to be critical for interaction with an inhibitor of Tn5 transposition. The inhibitor is a protein encoded by the same gene that encodes the transposase, but which is truncated at the N-terminal end of the protein, relative to the transposase. The approach of Weinreich et al. for determining the extent to which multimers are formed is suitable for determining whether a mutation falls within the scope of this element.

4.4.1.1.2.3. Reduced Inhibitory Activity Leading To Higher Levels Of Transposition

It is thought that when wild type Tn5 transposase multimerizes, its activity in trans is reduced. Presumably, a mutation in the dimerization region reduces or prevents multimerization, thereby reducing inhibitory activity and leading to levels of transposition 5- to 50-fold higher than are seen with the wild type transposase. The LP372 mutation achieves about 10-fold higher transposition levels than wild type. Likewise, other mutations (including mutations at a one or more amino acid) that reduce the ability of the transposase to multimerize would also function in the same manner as the single mutation at position 372, and would also be suitable in a transposase of the present invention. It may also be possible to reduce the ability of a Tn5 transposase to multimerize without altering the wild type sequence in the so-called dimerization region, for example by adding into the system another protein or non-protein agent that blocks the dimerization site. Alternatively, the dimerization region could be removed entirely from the transposase protein.

4.4.1.1.2.3.1. Inhibitor Protein

As was noted above, the inhibitor protein, encoded in partially overlapping sequence with the transposase, can interfere with transposase activity. As such, it is desired that the amount of inhibitor protein be reduced over the amount observed in wild type in vivo. For the present assay, the transposase is used in purified form, and it may be possible to separate the transposase from the inhibitor (for example, according to differences in size) before use. However, it is also possible to genetically eliminate the possibility of having any contaminating inhibitor protein present by removing its start codon from the gene that encodes the transposase.

An AUG in the wild type Tn5 transposase gene that encodes methionine at transposase amino acid 56 is the first codon of the inhibitor protein. However, it has already been shown that replacement of the methionine at position 56 has no apparent effect upon the transposase activity, but at the same time prevents translation of the inhibitor protein, thus resulting in a somewhat higher transposition rate. Weigand, T. W. and W. S. Reznikoff, "Characterization of Two Hypertransposing Tn5 Mutants," J. Bact. 174:1229-1239 (1992), incorporated herein by reference. In particular, the present inventors have replaced the methionine with an alanine in the preferred embodiment (and have replaced the

methionine-encoding AUG codon with an alanine-encoding GCC). A preferred transposase of the present invention therefore includes an amino acid other than methionine at amino acid position 56, although this change can be considered merely technically advantageous (since it ensures the absence of the inhibitor from the in vitro system) and not essential to the invention (since other means can be used to eliminate the inhibitor protein from the in vitro system).

4.4.1.2. Preferred Transposase Amino Acid Sequence

The most preferred transposase amino acid sequence known to the inventors differs from the wild type at amino acid positions 54, 56, and 372. The mutations at positions 54 and 372 separately contribute approximately a 10-fold increase to the rate of transposition reaction in vivo. When the mutations are combined using standard recombinant techniques into a single molecule containing both classes of mutations, reaction rates of at least about 100-fold higher than can be achieved using wild type transposase are observed when the products of the in vitro system are tested in vivo. The mutation at position 56 does not directly affect the transposase activity.

Other mutants from wild type that are contemplated to be likely to contribute to high transposase activity in vitro include, but are not limited to glutamic acid-to-lysine at position 110, and glutamic acid to lysine at position 345.

It is, of course, understood that other changes apart from these noted positions can be made to the modified transposase (or to a construct encoding the modified transposase) without adversely affecting the transposase activity. For example, it is well understood that a construct encoding such a transposase could include changes in the third position of codons such that the encoded amino acid does not differ from that described herein. In addition, certain codon changes have little or no functional effect upon the transposition activity of the encoded protein. Finally, other changes may be introduced which provide yet higher transposition activity in the encoded protein. It is also specifically envisioned that combinations of mutations can be combined to encode a modified transposase having even higher transposition activity than has been exemplified herein. All of these changes are within the scope of the present invention. It is noted, however, that a modified transposase containing the EK110 and EK345 mutations (both described by Weigand and

Reznikoff, supra, had lower transposase activity than a transposase containing either mutation alone.

4.4.1.3. Introduction Of Any Desired Transposable Element

After the enzyme is prepared and purified, as described supra, it can be used in the in vitro transposition reaction described above to introduce any desired transposable element from a donor DNA into a target DNA. The donor DNA can be circular or can be linear. If the donor DNA is linear, it is preferred that the repeat sequences flanking the transposable element should not be at the termini of the linear fragment but should rather include some DNA upstream and downstream from the region flanked by the repeat sequences.

The transposable element between the OE termini can include any desired nucleotide sequence. The length of the transposable element between the termini should be at least about 50 base pairs, although smaller inserts may work. No upper limit to the insert size is known. However, it is known that a donor DNA portion of about 300 nucleotides in length can function well. By way of non-limiting examples, the transposable element can include a coding region that encodes a detectable or selectable protein, with or without associated regulatory elements such as promoter, terminator, or the like.

If the element includes such a detectable or selectable coding region without a promoter, it will be possible to identify and map promoters in the target DNA that are uncovered by transposition of the coding region into a position downstream thereof, followed by analysis of the nucleic acid sequences upstream from the transposition site.

Likewise, the element can include a primer binding site that can be transposed into the target DNA, to facilitate sequencing methods or other methods that rely upon the use of primers distributed throughout the target genetic material. Similarly, the method can be used to introduce a desired restriction enzyme site or polylinker, or a site suitable for another type of recombination, such as a cre-lox, into the target.

The invention can be better understood upon consideration of the following examples which are intended to be exemplary and not limiting on the invention.

4.4.2. Generation of functional fusion protein products

The instant invention provides constructs and methods for the rapid and efficient generation of functional fusion protein products with either carboxy-terminal or amino-terminal fusions. Functional fusion proteins are those which retain some of the activity of the original domains, and/or those which have a newly created activity. Throughout this specification, reference is made to two types of fusions: carboxy terminal fusions and amino terminal fusions. In this text we use amino and carboxy terminal fusions to refer to the end of the domain inside of the Mu elements which is fused to the target molecule. Thus, carboxy terminal fusion elements are those with a protein domain inside of the Mu which extends out of the Mu element such that the exogenous protein is fused to the amino end of the endogenous protein. The amino terminal fusion elements are those that create fusions with a target gene extending into the element such that the exogenous protein is fused to the carboxy terminal of the endogenous protein.

An overview for generating both amino- and carboxy- terminal fusion proteins with the transposable elements is outlined. A domain of interest is inserted in one of two possible orientations into the end of a transposable element, such that a continuing open reading frame extends out through the element (for carboxy-terminal fusions) or in through the element (for amino-terminal fusions). Transposition into a target sequence allows random generation of hybrid proteins. Functional fusions can be selected or screened for.

4.4.2.1. Mini-Mu transposon elements & macro-Mu elements

The instant invention provides for mini-Mu transposon elements with convenient polylinker sites for inserting protein domains into the transposable element. The instant invention also provides for macro-Mu elements. The transposable elements of the instant invention have been designed to have a shorter transposon end which becomes incorporated into the fusion product while still retaining their high transposition frequencies. Unlike other elements used to date they can be used to make both amino and carboxy-terminal fusions since they have open reading frames extending in both directions. The examples below demonstrate the application of the teachings of the instant invention for using the new elements to create protein fusions with two fully active domains and their usefulness as a general tool for protein fusion. Other examples

demonstrate new features such as regulatable promoters incorporated into the transposable element, which can allow control of expression for promoterless domains or domains which may exhibit some degree of lethality to the cell. The constructs and methods described here allow high frequency random fusion of two domains at multiple sites so that the optimal fusion junctions can be selected.

4.4.2.2. The constructs

The constructs and teachings of the instant invention provide a powerful tool which will aid in the development of new enzymes for biocatalytic applications such as bioremediation and industrial biocatalysis, and for other industrial applications such as biosensors and strain development. Many different combinations of fusions between two domains can be generated rapidly and screened for activity. As an extension of protein evolution, this will be a powerful technique for production of novel chimeric proteins. In contrast to the difficulties inherent in a traditional rational protein design approach which employs traditional molecular biological techniques to generate fusion proteins, the instant invention provides for rapid and efficient "randomized experimentation" whereby the transposon fusion events are utilized to generate a panel of fusion proteins. From this panel of fusion proteins, selection for functionally expressed products, results in efficient screening for successful fusions.

While traditional experimentation with transposable elements resulted in the ablation of endogenous protein translation, and substitution of the translation and expression of the reporter gene for study of gene expression and regulation, the instant invention teaches modifications of these elements which allow for the generation of functional fusion proteins which can contain novel activities. The instant invention teaches for the first time the construction of transposable elements which can transpose exogenous DNA into target DNA in both orientations which can result in the functional expression of fusion proteins which have the exogenous protein linked either on the amino-end, or at the carboxy-end of the target protein.

The constructs and methods of the instant invention are not limited to the use of proteins with known crystal structures, but is capable of generating functional fusion proteins from

the randomized combination of functional domains. Instead of being limited to the mere tethering of functional domains, the constructs and methods of the instant invention allow for the fusion protein product to occur at randomized "truncated" sites of the target protein. This allows for the rapid generation of and screening for functional fusion proteins that are incorporated into a truncated form of the targeted protein.

Thus the instant invention has provided for creating new mini-Mu elements with polylinkers for domain insertion and open reading frames both into and out of the element. The elements have reduced ends while maintaining their high transposition frequency. The constructs of the instant invention and examples below show the ability of these elements to make carboxy-terminal protein fusions between a Protein A domain carried on the element and both selectable (chloramphenicol-acetyl transferase) and screenable (beta-galactosidase) proteins.

The constructs of the instant invention are useful for making amino-terminal fusions and demonstration of kinetic observations with beta-galactosidase showing that the K_m value is maintained even if it is fused to different locations.

4.4.2.3. In vivo settings for generating functional fusion proteins

While the constructs and methods of the instant invention are functional in an in vitro setting, a major advantage of the instant invention is the application of in vivo settings for generating functional fusion proteins. The development of an in vivo protein engineering system that is both versatile and easy to use will have a significant impact on the way proteins with specific activities are defined. This invention results in the development of a refined system for the in vivo engineering of proteins. The constructs and methods of the instant invention are powerful enhancements to the genetic tools available to the molecular biologist. The constructs of the instant invention can be provided as kits for the engineering of proteins.

One might imagine that an in vivo system may have problems since biological systems are limited to the properties of the particular system. One of the advantages of the Mu transposition system is that it is the most general and well defined of the transposition systems available (Symonds, Toussaint et al. (1987). Phage Mu.). The biological

properties of the system are advantages which make Mu simple to use. Mu provides very convenient methods for isolating independent insertion events by transducing them to a new cell. Creation of insertions requires only a temperature shift. We have favored the Mu transposition system from the start because an untrained technician may be taught to reproducibly use the Mu system in just a few days. While the initial design and construction of a Mu system can be difficult, the finished system's simplicity lends itself to commercialization. There is much to be said for in vitro systems. The in vivo system of the instant invention was intended to be an additional protein engineering tool to enhance the methods of producing novel active protein fusions when used in conjunction with basic in vitro methods, or with additional novel in vitro methods of the instant invention. There has been no previous demonstration of the altered transcription through the end of a Mu element. Specifically, transcription through the end of the Mu element might reduce transposition frequencies. Surprisingly we have found that the transcription of the Protein A gene through the end of Mu did not alter transposition frequencies.

The instant invention encompasses both the amino and carboxy-terminal fusion elements because they are useful in different applications. The carboxy-terminal element is more useful for making fusions with a cDNA which has been cloned from a eukaryotic organism, especially if one wants to ultimately express it in another organism. Amino-terminal elements are more useful for some applications because they can only be activated by a fusion event if they are missing the start codon. The instant specification teaches new Mu elements with promoters for expression in the experimental design. One of the main commercial applications of the system is to generate new enzymes for medical or industrial uses both in vivo and in vitro. E. coli will likely be the production organism if there are no post-translational modification problems associated with the enzymes. But it is possible to use any Mu compatible strain as the production organism.

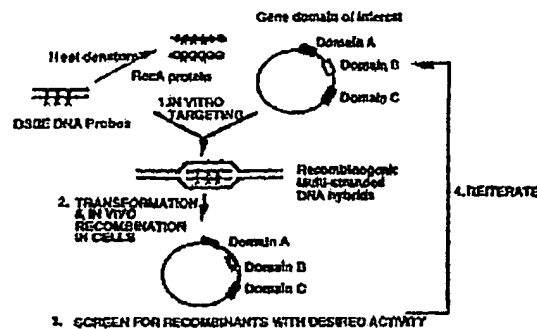
4.5. Additional Applications

It is envisioned that in addition to the uses specifically noted herein, other applications will be apparent to the skilled molecular biologist. In particular, methods for introducing desired mutations into prokaryotic or eukaryotic DNA are very desirable. For example, at present it is difficult to knock out a functional eukaryotic gene by homologous recombination with an inactive version of the gene that resides on a plasmid. The difficulty arises from the need to flank the gene on the plasmid with extensive upstream and downstream sequences. Using this system, however, an inactivating transposable element containing a selectable marker gene (e.g., neo) can be introduced in vitro into a plasmid that contains the gene that one desires to inactivate. After transposition, the products can be introduced into suitable host cells. Using standard selection means, one can recover only cell colonies that contain a plasmid having the transposable element. Such plasmids can be screened, for example by restriction analysis, to recover those that contain a disrupted gene. Such clones can then be introduced directly into eukaryotic cells for homologous recombination and selection using the same marker gene.

Also, one can use the system to readily insert a PCR-amplified DNA fragment into a vector, thus avoiding traditional cloning steps entirely. This can be accomplished by (1) providing suitable a pair of PCR primers containing OE termini adjacent to the sequence-specific parts of the primers, (2) performing standard PCR amplification of a desired nucleic acid fragment, (3) performing the in vitro transposition reaction of the present invention using the double-stranded products of PCR amplification as the donor DNA.

The invention is not intended to be limited to the foregoing examples, but to encompass all such modifications and variations as come within the scope of the appended claims.

5. Homologous Recombination



5.1. Homologous Recombination For The Generation Of Deletions And Insertions

The invention relates to compositions and methods of rapidly evolving specific protein domains using a library of nucleic acid filaments and a recombinase polypeptide or peptide. The invention relates to compositions and methods for targeting sequence modifications in one or more genes of a related family of genes using enhanced homologous recombination techniques. The invention also relates to compositions and methods for isolating and identifying novel members of homologous sequences families. These techniques may be used to create animal or plant models of disease as well as to identify new targets for drug or pathogen screening.

5.1.1. Evolution Of Genes

In nature, the evolution of genes and their encoded proteins occurs through an equilibrium between recombination or mutation and selection. While evolution in nature takes millions of years, in vitro methods and compositions have been developed to evolve proteins, with improved and novel functions, in a matter of hours to days.

5.1.1.1. Through Mutagenesis

Current in vitro gene evolution methods utilize repeated cycles of random mutagenesis or random nicking and mixing of related genes containing mutations in PCR-based random

recombination. These methods couple multiple rounds of in vitro mutagenesis with screening systems to produce and identify the desired mutants or recombinants (Stemmer 1994. Nature 370:389-391; Arnold 1996. Chemical Engineering Science 51:5091-5102). Research has shown, however, that the mutations of interest tend to occur in those regions or domains that are directly related to function (Chen and Arnold. 1993. PNAS USA 90:5618-5622).

However, these mutagenesis methods produce random mutations throughout the gene of interest which requires the need to screen large numbers of uninteresting or deleterious mutants. The labor-intensive and time consuming aspects of these methods are further complicated by the necessity of multiple rounds of subcloning and can be extremely challenging if the screening system is complex and does not utilize a selection system.

5.1.1.2. Homologous Recombination (HR)

Homologous recombination (HR) is defined as the exchange of homologous or similar DNA sequences between two DNA molecules. An essential feature of HR is that the enzymes responsible for the recombination event can pair any homologous sequences as substrates. The ability of HR to transfer genetic information between DNA molecules makes targeted homologous recombination a very powerful method in genetic engineering and gene manipulation. Both genetic and cytological studies have indicated that such a crossing-over process occurs between pairs of homologous chromosomes during meiosis in higher organisms.

5.1.1.3. Site-Specific Recombination

Alternatively, in site-specific recombination, exchange occurs at a specific site, as in the integration of phage λ into the E coli chromosome and the excision of lambda DNA from it. Site-specific recombination involves specific inverted repeat sequences; e.g. the Cre-loxP and FLP-FRT systems. Within these sequences there is only a short stretch of homology necessary for the recombination event, but not sufficient for it. The enzymes involved in this event generally cannot recombine other pairs of homologous (or nonhomologous) sequences, but act specifically.

5.1.1.4. Advantage Of Homologous Recombination Over Site-Specific Recombination

Although both site-specific recombination and homologous recombination are useful mechanisms for genetic engineering of DNA sequences, targeted homologous recombination provides a basis for targeting and altering essentially any desired sequence in a duplex DNA molecule, such as targeting a DNA sequence in a chromosome for replacement by another sequence. Site-specific recombination has been proposed as one method to integrate transfected DNA at chromosomal locations having specific recognition sites (O'Gorman et al. (1991) *Science* 251: 1351; Onouchi et al. (1991) *Nucleic Acids Res.* 19: 6373). Unfortunately, since this approach requires the presence of specific target sequences and recombinases, its utility for targeting recombination events at any particular chromosomal location is severely limited in comparison to targeted general recombination.

5.1.1.5. HR To Create Transgenic Plants, Animals, And Organisms

Homologous recombination has also been used to create transgenic plants and animals. Transgenic organisms contain stably integrated copies of genes or gene constructs derived from another species in the chromosome of the transgenic organism. In addition, gene targeted animals can be generated by introducing cloned DNA constructs of the foreign genes into totipotent cells by a variety of methods, including homologous recombination. For example, animals that develop from genetically altered totipotent cells can contain the foreign gene in all somatic cells and also in germ-line cells.

5.1.1.5.1. Current Methods Using Embryonic Stem Cells

Currently methods for producing transgenic and targeted animals have been performed on totipotent embryonic stem cells (ES) and with fertilized zygotes. ES cells have an advantage in that large numbers of cells can be manipulated easily by homologous recombination in vitro before they are used to generate targeted animals. Currently, however, only embryonic stem cells from mice have been shown to contribute to the germ line. Alternatively, DNA can also be introduced into fertilized oocytes by micro-injection into pronuclei which are then transferred into the uterus of a pseudo-pregnant recipient animal to develop to term. The ability of mammalian and human cells to incorporate exogenous genetic material into genes residing on chromosomes has demonstrated that these cells have the general enzymatic machinery for carrying out homologous

recombination required between resident and introduced sequences. These targeted recombination events can be used to correct mutations at known sites, replace genes or gene segments with defective ones, or introduce foreign genes into cells.

5.1.1.5.2. Frequency And Efficiency Of HR

HR can be used to add subtle mutations at known sites, replace wild type genes or gene segments or introduce completely foreign genes into cells. However, HR efficiency is very low in living cells and is dependent on several parameters, including the method of DNA delivery, how it is packaged, its size and conformation, DNA length and position of sequences homologous to the target, and the efficiency of hybridization and recombination at chromosomal sites. These variables severely limit the use of conventional HR approaches for gene evolution in cell based systems. (Kucherlapati et al. , 1984. PNAS; USA 81:3153-- 3157; Smithies et al. 1985. Nature 317:230-234; Song et al. 1987. PNAS USA 84:6820-6824; Doetschman et al. 1987. Nature 330:576-578; Kim and Smithies. 1988. Nuc. Acids. Res. 16:8887- 8903; Koller and Smithies. 1989. PNAS USA 86:8932-8935; Shesely et al. 1991. PNAS USA 88:4294- 4298; Kim et al. 1991. Gene 103:227-233).

5.1.1.5.2.1. Enhancement By The Presence Of Recombinase Activities

The frequency of HR is significantly enhanced by the presence of recombinase activities in cellular and cell free systems. Several proteins or purified extracts that promote HR (i.e., recombinase activity) have been identified in prokaryotes and eukaryotes (Cox and Lehman., 1987. Annu. Rev. Biochem. 56:229-262; Radding. 1982. Annual Review of Genetics 16:405-547; McCarthy et al. 1988. PNAS; USA 85:5854-5858). These recombinases promote one or more steps in the formation of homologously-paired intermediates, strand-exchange, and/or other steps.

Recent advances have resulted in techniques allowing enhanced homologous recombination (EHR) using recombinases such as recA and Rad51 and single-stranded nucleic acids that have sequence heterologies. This allows sequence modifications to be specifically targeted to virtually any genomic position. See for example, PCT US93/03868 and PCT US98/05223, both of which are expressly incorporated herein by reference.

5.1.1.5.2.1.1. Recombinase Rec A: A Bacterial Protein That Catalyses Homologous Pairing And Strand Exchange Between Two Homologous DNA Molecules

The most studied recombinase to date is the RecA recombinase of *E. coli*, which is involved in homology search and strand exchange reactions (Cox and Lehman, 1987, *supra*). The bacterial RecA protein (Mr 37,842) catalyses homologous pairing and strand exchange between two homologous DNA molecules (Kowalczykowski et al. 1994. *Microbiol. Rev.* 58:401-465; West. 1992. *Annu. Rev. Biochem.* 61:603-640); Roca and Cox. 1990. *CRC Crit. Rev. Biochem. Mol. Biol.* :415-455; Radding. 1989. *Biochim. Biophys. Acta.* 1008:131-145; Smith. 1989. *Cell* 58:807-809).

RecA protein binds cooperatively to any given sequence of single-stranded DNA with a stoichiometry of one RecA protein monomer for every three to four nucleotides in DNA (Cox and Lehman, 1987, *supra*). This forms unique right handed helical nucleoprotein filaments in which the DNA is extended by 1.5 times its usual length (Yu and Egelman 1992. *J. Mol. Biol.* 227:334-346). These nucleoprotein filaments, which are referred to as DNA probes, are crucial "homology search engines" which catalyze DNA pairing. Once the filament finds its homologous target gene sequence, the DNA probe strand invades the target and forms a hybrid DNA structure, referred to as a joint molecule or D-loop (DNA displacement loop) (McEntee et al. 1979. *PNAS USA* 76:2615-2619; Shibata et al. 1979. *PNAS USA* 76:1638-1642). The phosphate backbone of DNA inside the RecA nucleoprotein filaments is protected against digestion by phosphodiesterases and nucleases.

RecA protein is the prototype of a universal class of recombinase enzymes which promote probe-target pairing reactions. Recently, genes homologous to *E. coli* RecA (the Rad51 family of proteins) were isolated from all groups of eukaryotes, including yeast and humans. Rad51 protein promotes homologous pairing and strand invasion and exchange between homologous DNA molecules in a similar manner to RecA protein (Sung. 1994. *Science* 265:1241-1243; Sung and Robberson. 1995. *Cell* 82:453-461; Gupta et al. 1997. *PNAS USA* 94:463-468; Baumann et al. 1996. *Cell* 87:757-766).

5.1.1.5.3. Functional Genomics: The Correlation Of Genotype And Phenotype

One area of pressing interest in biology is within the area of "functional genomics", i.e. the correlation of genotype and phenotype. This requires animal systems, since phenotypic changes must be evaluated in vivo. Similarly, and related to this idea, is the elucidation and characterization of gene families, i.e. genes or proteins that are structurally related, i.e. they have sequence homologies between the members of the family. Since presumably many, if not most, disease states are caused by multiple gene interactions, the ability to evaluate interactions among genes, and particularly within or between gene families, at the phenotype level, would be extremely valuable.

The functional genomics tools that allow facile identification and engineering of gene family members in animals and cells, however, are not yet available. While the amino acid sequence motifs shared between gene family members may be identical, due to degeneracy in the DNA code, the DNA sequence identity may be significantly less. Hence, one criterion necessary for genetic modifications of gene family members is development of homologous recombination technologies that can be used to clone and modify similar DNA sequences that share little sequence identity. This is particularly important since homologous recombination in cells normally requires significant sequence identity to work efficiently. Relaxing the amount of sequence identity needed for homologous recombination allows greater flexibility to target related genes for creating transgenic animals and cells containing modifications in gene family consensus sequences, and also will allow the rapid cloning, generation of gene family specific libraries, and evolution of gene family members. Accordingly, it is an object of the invention to provide an efficient method of domain specific gene evolution that generates maximal diversity but increases the probability of identifying a gene of interest.

5.2. Domain Specific Gene Evolution

5.2.1.1. Domain Specific Gene Evolution - Comprising Forming A Plurality Of Recombination Intermediates Comprising A Target Nucleic Acid Encoding An Amino Acid Sequence Of Interest, A Recombinase And A Plurality Of Targeting Polynucleotides

The present invention provides methods of domain specific gene evolution comprising forming a plurality of recombination intermediates comprising a target nucleic acid encoding an amino acid sequence of interest, a recombinase and a plurality of targeting polynucleotides. The targeting polynucleotides are substantially complementary to each other and each comprises a homology clamp that substantially correspond to or is substantially complementary to a predetermined sequence of the target nucleic acid and comprise random or degenerate sequences. The predetermined sequence encodes a domain of the amino acid sequence. The method further comprises contacting the intermediate with a recombination proficient cell, whereby a library of altered target nucleic acids are produced. The altered target nucleic acids are expressed in the cell to generate a pool of variant amino acid sequences. The method further comprises selecting and isolating a cell comprising an altered target nucleic acid that expressed a variant amino acid having a desired activity.

5.2.1.2. Comprising Forming A Recombination Intermediate Comprising A Target Nucleic Acid Encoding An Amino Acid Sequence Of Interest, A Recombinase And A Pair Of Targeting Poly Nucleotides

In another aspect of the invention, a method of domain specific gene evolution comprises forming a recombination intermediate comprising a target nucleic acid encoding an amino acid sequence of interest, a recombinase and a pair of targeting polynucleotides. The targeting polynucleotides are substantially complementary to each other and each comprises a homology clamp that substantially corresponds to or is substantially complementary to a predetermined sequence of the target nucleic acid. The predetermined sequence encodes a domain of the amino acid sequence. The method further comprises contacting the intermediate with a single-strand specific nuclease or junction-specific nuclease to form a nicked or open-ended target nucleic acid. The regions adjacent to the

hybridized region or junctions are susceptible to nucleases. The target nucleic acid is reassembled and recombined to produce a library of altered target nucleic acids. The target nucleic acids are expressed to generate a pool of variant amino acid sequences. The variant amino acid sequences are selected and characterized to identify an altered target nucleic acid encoding a variant amino acid sequence of interest.

In a further aspect, each method is repeated one or more times to further evolve a variant amino acid sequence having a desired activity. In yet another aspect, more than one domain or a protein is evolved simultaneously.

5.2.1.3. Compositions

It is an object of the present invention to provide compositions comprising at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for a gene family.

In an additional aspect, the invention provides compositions comprising at least one recombinase and a plurality of pairs of single stranded targeting polynucleotides, where the plurality of pairs comprises a set of degenerate probes encoding the consensus sequence.

In a further aspect, the invention provides kits comprising the compositions of the invention and at least one reagent.

5.2.1.4. Methods For Targeting A Sequence Modification In At Least One Member Of A Consensus Family Of Genes In A Cell By Homologous Recombination.

In an additional aspect, the invention provides methods for targeting a sequence modification in at least one member of a consensus family of genes in a cell by homologous recombination. The method comprises introducing into at least one cell at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for the family. The method can additionally comprise identifying a target cell having a targeted sequence modification.

5.2.1.4.1. Methods Of Making A Non- Human Organism With A Targeted Sequence Modification In At Least One Member Of A Gene Family

In a further aspect, the invention provides methods of making a non- human organism with a targeted sequence modification in at least one member of a gene family. The method comprises introducing into a cell at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for said family. The cell is then subjected to conditions that result in the formation of an animal, and the animal has at least one modification in at least one member of a consensus family of genes.

In a further aspect, the invention provides non-human organisms containing a sequence modification in an endogenous consensus functional domain of a gene member of a gene family.

5.2.1.5. Methods Of Isolating A Member Of A Gene Family Comprising A Protein Consensus Sequence

In an additional aspect, the invention provides methods of isolating a member of a gene family comprising a protein consensus sequence. The method comprises adding to a complex mixture of nucleic acids at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for said family. At least one of the targeting polynucleotides comprises a purification tag. The method is done under conditions whereby the targeting polynucleotides form a complex with the member, and the family member is isolated using said purification tag. The complex nucleic acid mixture may be a cDNA library, a cell, RNA or a restriction endonucleases genomic digest.

5.3. Targeting A Predetermined Nucleic Acid Sequence That Encodes A Specific Protein Domain, To Make A Plurality Of Targeted Sequence Modifications

The present invention provides methods and compositions for domain specific gene evolution. In one aspect of the invention, the method comprises targeting a predetermined nucleic acid sequence that encodes a specific protein domain, to make a plurality of targeted sequence modifications. That is, by targeting the recombinogenic probes of the

invention to particular protein domains, gene evolution and selection are targeted to specific domains known or believed to harbor specific activities or functions. These methods create maximal diversity in specific domains of interest, thereby, decreasing the size of the library of mutations that are to be screened and increasing the probability of finding a gene with improved or desired attributes. Therefore, the libraries of the present invention are enriched for advantageous or interesting mutations or recombinant sequence(s).

5.3.1. Combining A Plurality Of Pairs Of Single-Stranded Targeting Poly Nucleotides, A Predetermined Target Nucleic Acid, And A Recombinase To Form A Polynucleotide:Target Nucleic Acid Complex

Accordingly, the methods comprise combining a plurality of pairs of single-stranded targeting poly nucleotides, a predetermined target nucleic acid, and a recombinase to form a polynucleotide:target nucleic acid complex. The targeting polynucleotides comprise at least one homology clamp for targeting a predetermined domain of a target nucleic acid and randomized or degenerate sequences. The complex is optionally introduced into a plurality of recombination proficient cells which catalyze strand exchange and homologous recombination intracellularly to produce a library of modified nucleic acids. Cells are selected and isolated that comprise a modified nucleic acid that encodes a polypeptide having a desired property. The process is preferably repeated iteratively to further evolve the target domain of interest.

5.3.2. Domain Specific DNA Nicking

In another aspect of the invention, methods of domain specific DNA nicking are provided for domain specific gene evolution. This method comprises combining a pair of single-stranded targeting polynucleotides, a predetermined target nucleic acid, and a recombinase to form a polynucleotide:target nucleic acid complex. The targeting polynucleotides are substantially complementary and comprise at least one homology clamp for targeting a predetermined domain of a target nucleic acid. The polynucleotide:target nucleic acid complex is treated with a single-strand specific nuclease, which preferentially nicks the regions flanking the polynucleotide:target nucleic acid complex region (Ferrin and Camerini-Otero. 1991. Science. 254 1494-1497). That is, the domain is protected from recombination by the initial presence of the recombinase in the complex. The nuclease is

inactivated and the complex dissociated. The nicked target nucleic acid is reassembled and recombined by PCR to produce a library of nucleic acids with preferential modifications in the nicked regions. The library of modified nucleic acids can be introduced into a host cell and expressed. Cells are selected and isolated that comprise a modified nucleic acid that encodes a polypeptide having a desired property. This process is repeated iteratively to further evolve the predetermined targeted domain of interest.

In each of the methods described above, single domains and optionally multiple domains are targeted. The methods and compositions described above are optionally used in combination for domain specific gene evolution. For example, individual or multiple rounds of domain specific DNA nicking are followed or interspersed with one or more rounds of domain specific evolution employing a plurality of targeting polynucleotides described above.

The methods of the present invention also avoid multiple subcloning steps. This is particularly relevant when large complex vectors such as lambda, BACS, PACS, YACS, MACS and other genomic DNAs are used and where multiple subcloning steps make mutagenesis and shuffling of unique sites in large vectors particularly tedious and time consuming.

5.3.3. Generating Homologous Recombination Intermediates In Vitro, Panels Or Libraries Of Mutagenized And Shuffled Genes To Generate In Vitro Evolution

Accordingly, the present invention provides methods to introduce recombinogenic probe or hybrid complexes into recombination proficient cells to link in vitro and in vivo recombination and evolution processes. By generating homologous recombination intermediates in vitro, panels or libraries of mutagenized and shuffled genes are generated for in vitro evolution. The link to in vivo systems allows in vivo selection of evolved genes encoding proteins of a desired characteristic. The present invention can thus be used in a variety of important ways.

5.3.3.1. Methods Can Be Used In The Creation Of Transgenic Organisms, Animal, And Plant Models Of Disease

First, these methods can be used in the creation of transgenic organisms, animal, and plant models of disease. Thus, for example, domain-specific targeting polynucleotides used in homologous recombination methods can generate animals that have a wide variety of mutations in a wide variety of functionally related genes, potentially resulting in a wide variety of phenotypes, including phenotypes related to disease states. This may also be done on a cellular level, to identify genes involved in cellular phenotypes, i.e. target identification.

5.3.3.2. Identify "Reversion" Genes, Genes That Can Modulate Disease States

Secondly, domain targeting can be used in cells or animals that are diseased or altered; in essence, domain targeting can be done to identify "reversion" genes, genes that can modulate disease states caused by different genes, either genes within the same gene family or a completely different gene family. Thus, for example the loss of one type of enzymatic activity, resulting in a disease phenotype, may be compensated by alterations in a different but homologous enzymatic activity.

5.3.3.3. Creation Of Libraries Of Altered Nucleic Acids

In addition, the methods may be used in the creation of libraries of altered nucleic acids, including extrachromosomal sequences, and can be expressed in cells to produce libraries of altered proteins, which then can be screened for any number of useful or interesting properties, including, but not limited to, increased or altered stability (thermal, pH,

oxidants, to proteases, etc.); altered specificity (for example, in the case of enzymes); altered binding; modified activity and other desirable properties, such as, altered immunogenicity.

5.3.4. Use Of Homology Motif Tags (HMTs) In Targeted Homologous Recombination To Elucidate Disease Mechanisms And To Identify Disease Targets Contained Within Gene Families

The present invention is directed to the use of homology motif tags (HMTs) in targeted homologous recombination to elucidate disease mechanisms and to identify disease targets contained within gene families related by the presence of one or more common domains. That is, there are a large number of gene families that contain genes related by the presence of similar functional domains, i.e. binding domains for substrates or other proteins, enzymatic domains such as kinase or protease domains, signaling and regulator domains, receptor binding domains, ATP binding domains, leucine zipper domains, zinc finger domains, etc. These functional domains frequently result in primary sequence homology; that is, related functional domains have related sequences. Many of these functional domains have been studied and so-called "consensus sequences" identified; that is, an average sequence derived from a number of related sequences. Each residue (or set of residues) of the consensus sequence is the most frequent at that position in the set under consideration. Consensus sequences can be either amino acid or nucleic acid consensus sequences, with amino acid sequences being used to generate nucleic acid consensus sequences.

Interestingly, while a wide variety of gene families are known, the majority of drug targets come from only four of these gene families. These are the G-protein coupled or seven-transmembrane domain receptors, nuclear (hormone) receptors, ion channels, esterases. Other important gene families are enzymes, including recombinases. Of the top 100 pharmaceutical drugs, 18 bind to seven- transmembrane receptors, 10 to nuclear receptors and 16 to ion channels.

By using HMTs directed to the consensus sequences of gene families for homologous recombination and particularly enhanced homologous recombination methods, sequence modifications may be made to any number of targeted genes in a related family.

5.3.4.1. Methods And Compositions Utilizing Homology Motif Tags (HMTs) Or Consensus Sequences

Accordingly, the present invention provides methods and compositions utilizing homology motif tags (HMTs) or consensus sequences. By "homology motif tag" or "protein consensus sequence" herein is meant an amino acid consensus sequence of a gene family. By "consensus nucleic acid sequence" herein is meant a nucleic acid that encodes a consensus protein sequence of a functional domain of a gene family. In addition, "consensus nucleic acid sequence" can also refer to cis sequences that are non-coding but can serve a regulatory or other role. As outlined below, generally a library of consensus nucleic acid sequences are used, that comprises a set of degenerate nucleic acids encoding the protein consensus sequence. A wide variety of protein consensus sequences for a number of gene families are known. A "gene family" therefore is a set of genes that encode proteins that contain a functional is domain for which a consensus sequence can be identified. However, in some instances, a gene family includes non-coding sequences; for example, consensus regulatory regions can be identified. For example, gene family/consensus sequences pairs are known for the G- protein coupled receptor family, the AAA-protein family, the bZIP transcription factor family, the mutS family, the recA family, the Rad51 family, the dmel family, the recF family, the SH2 domain family, the Bcl- 2 family, the single-stranded binding protein family, the TFIID transcription family, the TGF-beta family, the TNF family, the XPA family, the XPG family, actin binding proteins, bromodomain GDP exchange factors, MCM family, ser/thr phosphatase family, etc.

As will be appreciated by those in the art, the proteins of the gene families generally do not contain the exact consensus sequences; generally consensus sequences are artificial sequences that represent the best comparison of a variety of sequences. The actual sequence that corresponds to the functional sequence within a particular protein is termed a "consensus functional domain" herein; that is, a consensus functional domain is the actual sequence within a protein that corresponds to the consensus sequence. A consensus functional domain may also be a "predetermined endogenous DNA sequence" (also referred to herein as a "predetermined target sequence") that is a polynucleotide sequence contained in a target cell. Such sequences can include, for example, chromosomal

sequences (e.g., structural genes, regulatory sequences including promoters and enhancers, recombinatorial hotspots, repeat sequences, integrated proviral sequences, hairpins, palindromes), episomal or extrachromosomal sequences (e.g., replicable plasmids or viral replication intermediates) including chloroplast and mitochondrial DNA sequences. By "predetermined" or "pre-selected" it is meant that the consensus functional domain target sequence may be selected at the discretion of the practitioner on the basis of known or predicted sequence information, and is not constrained to specific sites recognized by certain site-specific recombinases (e.g., FLP recombinase or CRE recombinase). In some embodiments, the predetermined endogenous DNA target sequence will be other than a naturally occurring germline DNA sequence (e.g., a transgene, parasitic, mycoplasmal or viral sequence).

5.3.4.1.1. Gene Family Is The G-Protein Coupled Receptor Family

5.3.4.1.1.1. Subfamily 1 Also Called R7G Proteins

In a preferred embodiment, the gene family is the G-protein coupled receptor family, which has over 900 identified members, including several subfamilies. In a preferred embodiment, the G-protein coupled receptors are from subfamily 1 and are also called R7G proteins. They are an extensive group of receptors which recognize hormones, neurotransmitters, odorants and light and transduce extracellular signals by interaction with guanine (G) nucleotide-binding proteins. The structure of all these receptors is thought to be virtually identical, and they contain seven hydrophobic regions, each of which putatively spans the membrane. The N-terminus is extracellular and is frequently glycosylated, and the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three cytoplasmic loops to link the seven transmembrane regions. G- protein coupled receptors include, but are not limited to: the class A rhodopsin first subfamily, including amine (acetylcholine (muscarinic), adrenoceptors, dopamine, histamine, serotonin, octopamine), peptides (angiotensin, bombesin, bradykinin, C5a anaphylatoxin, Fmet-leu-phe, interleukin-8, chemokine, CCK, endothelin, melanocortin, neuropeptide Y, neurotensin, opioid, somatostatin, tachykinin, thrombin, vasopressin-like, galanin, proteinase activated), hormone proteins (follicle stimulating hormone, luteinizing hormone-releasing hormone, thyrotropin), rhodopsin (vertebrate), olfactory (olfactory type 1-11, gustatory), prostanoid (prostaglandin, prostacyclin, thromboxane), nucleotide (adenosine, purinoceptors), cannabis, platelet activating factor, gonadotropin-releasing

hormone (gonadotropin releasing hormone, thyrotropin-releasing hormone, growth hormone secretagogue), melatonin, viral proteins, MHC receptor, Mas proto-oncogene, EBV-induced and glucocorticoid induced; the class B secretin second subfamily, including calcitonin, corticotropin releasing factor, gastric inhibitory peptide, glucagon, growth hormone releasing hormone, parathyroid hormone, secretin, vasoactive intestinal polypeptide, and diuretic hormone; the class C metabotropic glutamate third subfamily, including metabotropic glutamate and extracellular calcium-sensing agents; and the class D pheromone fourth subfamily. Because of the large number of family members, these large classes of GPCRs can be further subdivided into subfamilies. Examples of these subfamilies are calcitonin, glucagon, vasoactive and parathyroid are from class B; and acetylcholine, histamine angiotensin, alpha2- and beta-adrenergic are from class A. From each subfamily small protein consensus sequences can be derived from sequence alignments. For example, there are 6 motifs for the metabotropic glutamate like GPCRs derived from the indicated number of family members. Using the protein consensus sequence, degenerate nucleic acid probes are made to encode the protein consensus sequence, as is well known in the art. The protein sequence is encoded by DNA triplets which are deduced using standard tables. In some cases additional degeneracy is used to enable production in one oligonucleotide synthesis. In many cases motifs were chosen to minimize degeneracy. Amplification of neighboring sequences can utilize two motifs as indicated by faithful or error prone amplification. Alternatively outside sequences can be used as is indicated using vector sequence. In addition degenerate oligos can be synthesized and used directly in the procedure without amplification. Double stranded (ds) DNA probes are denatured and coated with RecA or another recombinase such as Rad51. This material can be used to bind to and allow capture of specific clones from cDNA or genomic libraries. Alternatively this material can be introduced into cells producing transgenic cells or animals with alterations in related family members.

5.3.4.1.1.2. Second Subfamily Encoding Receptors That Bind Peptide Hormones That Do Not Show Sequence Similarity To The First R7G Subfamily

In addition to the first subfamily of G-protein coupled receptors, there is a second subfamily encoding receptors that bind peptide hormones that do not show sequence similarity to the first R7G subfamily. All the characterized receptors in this subfamily are coupled to G- proteins that activate both adenylyl cyclase and the phosphatidylinositol-

calcium pathway. However, they are structurally similar; like classical R7G proteins they putatively contain seven transmembrane regions, a glycosylated extracellular N-terminus and a cytoplasmic C-terminus. Known receptors in this subfamily are encoded on multiple exons, and several of these genes are alternatively spliced to yield functionally distinct products. The N-terminus contains five conserved cysteine residues putatively important in disulfide bonds. Known G-protein coupled receptors in this subfamily are listed above.

5.3.4.1.1.3. Third Subfamily Encoding Receptors That Bind Glutamate And Calcium But Do Not Show Sequence Similarity To Either Of The Other Subfamilies

In addition to the first and second subfamilies of G-protein coupled receptors, there is a third subfamily encoding receptors that bind glutamate and calcium but do not show sequence similarity to either of the other subfamilies. Structurally, this subfamily has signal sequences, very large hydrophobic extracellular regions of about 540 to 600 amino acids that contain 17 conserved cysteines (putatively involved in disulfides), a region of about 250 residues that appear to contain seven transmembrane domains, and a C-terminal cytoplasmic domain of variable length (50 to 350 residues). Known G- protein coupled receptors of this subfamily are listed above.

5.3.4.1.2. Gene Family Is The bZIP Transcription Factor Family

In a preferred embodiment, the gene family is the bZIP transcription factor family. This eukaryotic gene family encodes DNA binding transcription factors that contain a basic region that mediates sequence specific DNA binding, and a leucine zipper, required for dimerization. The bZIP family includes, but is not limited to, AP-1, ATF, CREB, CREM, FOS, FRA, GBF, GCN4, HBP, JUN, MET4, OCS1, OP, TAF1, XBP1, and YBBO.

In a preferred embodiment, the gene family is involved in DNA mismatch repair, such as mutL, hexB and PMS1. Members of this family include, but are not limited to, MLH1, PMS1, PMS2, HexB and Mull. The protein consensus sequence is G-F-R-G-E-A-L.

In a preferred embodiment, the gene family is the mutS family, also involved in mismatch repair of DNA, directed to the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. MutS gene family members include, but are not limited to, MSH2, MSH3, MSH6 and MutS. In a preferred embodiment, the gene family is the recA family. The bacterial recA is essential for homologous recombination and recombinatorial repair of DNA damage. RecA has many

activities, including the formation of nucleoprotein filaments, binding to single stranded and double stranded DNA, binding and hydrolyzing ATP, recombinase activity and interaction with *lexA* causing *lexA* activation and autocatalytic cleavage. RecA family members include those from *E. coli*, *drosophila*, human, lily, etc. specifically including but not limited to, *E. coli* *recA*, *Rec1*, *Rec2*, *Rad51*, *Rad51 B*, *Rad51 C*, *Rad51 D*, *Rad51 E*, *XRCC2* and *DMC1*.

5.3.4.1.3. Gene Family Is The RecF Family

In a preferred embodiment, the gene family is the *recF* family. The prokaryotic *recF* protein is a single- stranded DNA binding protein which also putatively binds ATP. *RecF* is involved in DNA metabolism; it is required for recombinatorial DNA repair and for induction of the SOS response. *RecF* is a protein of about 350 to 370 amino acid residues; there is a conserved ATP-binding site motif "A" in the N-terminal section of the protein as well as two other conserved regions, one located in the central section and the other in the C-terminal section.

5.3.4.1.4. Gene Family Is The Bcl-2 Family

In a preferred embodiment, the gene family is the Bcl-2 family. Programmed cell death (PCD), or apoptosis, is induced by events such as growth factor withdrawal and toxins. It is generally controlled by regulators, which have either an inhibitory effect (i.e. anti-apoptotic) or block the protective effect of inhibitors (pro-apoptotic). Many viruses have found a way of countering defensive apoptosis by encoding their own anti-apoptotic genes thereby preventing their target cells from dying too soon.

All proteins belonging to the Bcl-2 family contain at least one of a BH1, BH2, BH3 or BH4 domain. All anti-apoptotic proteins contain BH1 and BH2 domains, some of them contain an additional N-terminal BH4 domain (such as Bcl-2, Bcl-x(L), Bcl-W, etc.), which is generally not found in pro-apoptotic proteins (with the exception of Bcl-x(S). Generally all pro-apoptotic proteins contain a BH3 domain (except for Bad), thought to be crucial for the dimerization of the proteins with other Bcl-2 family members and crucial for their killing activity. In addition, some of the pro- apoptotic proteins contain BH1 and BH2 domains (such as Bax and Bak). The BH3 domain is also present in some anti-

apoptosis proteins, such as Bcl-2 and Bcl-x(L). Known Bcl-2 proteins include, but are not limited to, Bcl-2, Bcl-x(L), Bcl-W, Bcl-x(S), Bad, Bax, and Bak.

5.3.4.1.5. Gene Family Is The Site-Specific Recombinase Family

In a preferred embodiment, the gene family is the site-specific recombinase family. Site-specific recombination plays an important role in DNA rearrangement in prokaryotic organisms. Two types of site-specific recombination are known to occur: a) recombination between inverted repeats resulting in the reversal of a DNA segment; and b) recombination between repeat sequences on two DNA molecules resulting in their coinTEGRATION, or between repeats on one DNA molecule resulting the excision of a DNA fragment. Site-specific recombination is characterized by a strand exchange mechanism that requires no DNA synthesis or high energy cofactor; the phosphodiester bond energy is conserved in a phospho-protein linkage during strand cleavage and re-ligation.

Two unrelated families of recombinases are currently known. The first, called the "phage integrase" family, groups a number of bacterial, phage and yeast plasmid enzymes. The second, called the "resolvase" family, groups enzymes which share the following structural characteristics: an N-terminal catalytic and dimerization domain that contains a conserved serine residue involved in the transient covalent attachment to DNA, and a C-terminal helix-turn-helix DNA-binding domain.

5.3.4.1.6. Gene Family Is The Single-Stranded Binding Protein Family

In a preferred embodiment, the gene family is the single-stranded binding protein family. The E coli single-stranded binding protein (ssb), also known as the helix-destabilizing protein, is a protein of 177 amino acids. It binds tightly as a homotetramer to a single-stranded DNA (ss-DNA) and plays an important role in DNA replication, recombination and repair. Members of the ssb family include, but are not limited to, E. coli ssb and eukaryotic RPA proteins.

5.3.4.1.7. Gene Family Is The TFIID Transcription Family

In a preferred embodiment, the gene family is the TFIID transcription family.

Transcription factor TRID (or TATA-binding protein, TBP), is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II.

TRID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources.

This region is necessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region.

5.3.4.1.8. Gene Family Is The TGF-beta Family

In a preferred embodiment, the gene family is the TGF-beta family. Transforming growth factor-beta (TGF-beta) is a multifunctional protein that controls proliferation, differentiation and other functions in many cell types. TGF-beta-1 is a protein of 112 amino acid residues derived by proteolytic cleavage from the C-terminal portion of the precursor protein. Members of the TGF-beta family include, but are not limited to, the TGF-1-3 subfamily (including TGF1, TGF2, and TGF3); the BMP3 subfamily (BM3B, BMP3); the BMP5-8 subfamily (BM8A, BMP5, BMP6, BMP7, and BMP8); and the BMP 2 & 4 subfamily (BMP2, BMP4, DECA).

In a preferred embodiment, the gene family is the TNF family. A number of cytokines can be grouped into a family on the basis of amino acid sequence, as well as structural and functional similarities. These include (1) tumor necrosis factor (TNF), also known as cachectin or TNF-alpha, which is a cytokine with a wide variety of functions. TNF-alpha can cause cytolysis of certain tumor cell lines; it is involved in the induction of cachexia; it is a potent pyrogen, causing fever by direct action or by stimulation of interleukin-1 secretion; and it can stimulate cell proliferation and induce cell differentiation under certain conditions; (2) lymphotoxin-alpha (LT-alpha) and lymphotoxin-beta (LT-beta), two related cytokines produced by lymphocytes and which are cytotoxic for a wide range of tumor cells in vitro and in vivo; (3) T cell antigen gp39 (CD40L), a cytokine that seems to be important in B-cell development and activation; (4) CD27L, a cytokine that plays a role in T-cell activation; it induces the proliferation of costimulated T cells and enhances

the generation of cytolytic T cells; (5) CD30L, a cytokine that induces proliferation of T-cells; (6) FASL, a cytokine involved in cell death; (8) 4-1 BBL, an inducible T cell surface molecule that contributes to T-cell stimulation; (9) OX40L, a cytokine that co-stimulates T cell proliferation and cytokine production; and (10), TNF-related apoptosis inducing ligand (TRAIL), a cytokine that induces apoptosis.

5.3.4.1.9. Gene Family Is The XPA Family

In a preferred embodiment, the gene family is the XPA family. Xeroderma pigmentosa (XP) is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. Skin cells associated with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of 7 genetic complementation groups involved in this disorder: XPA to XPG. XPA is the most common form of the disease and is due to defects in a 30 kD nuclear protein called XPA or (XPAC). The sequence of XPA is conserved from higher eukaryotes to yeast (gene RAD14). XPA is a hydrophilic protein of 247 to 296 amino acid residues that has a C4- type zinc finger motif in its central section.

5.3.4.1.10. Gene Family Is The XPG Family

In a preferred embodiment, the gene family is the XPG family. The defect in XPG can be corrected by a 133 kD nuclear protein called XPG (or XPGC). Members of the XPG family include, but are not limited to, FEN1, XPG, RAD2, EXO1, and DIN7.

Once having identified a gene family and a consensus sequence, the compositions of the invention can be made. The compositions of the invention comprise at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each have a consensus homology clamp for a gene family.

5.3.5. Homologous Recombination

Accordingly, the present invention provides methods of homologous recombination. By "homologous recombination" (HR) herein is meant an exchange of homologous or similar DNA sequence between two DNA molecules. An essential feature of HR is that the enzyme responsible for the recombination event can pair any homologous sequences as

substrates. The ability of HR to transfer genetic information between DNA molecules makes targeted homologous recombination a very powerful method in genetic engineering and gene manipulation. HR can be used to insert, delete, and/or substitute any one or more nucleotides in a gene or gene segment or to introduce or delete genes in a targeted nucleic acid.

Once having identified a protein domain, the compositions of the invention can be made. The compositions of the invention comprise at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each have a domain homology clamp.

5.3.6. Recombinase

By "recombinase" herein is meant a protein or peptide (e.g. L2 peptide) that, when included with an exogenous targeting polynucleotide, provide a measurable increase in the recombination frequency and/or localization frequency between the targeting polynucleotide and an endogenous predetermined DNA sequence. Thus, in a preferred embodiment, increases in recombination frequency from the normal range of 10^{-8} to 10^{-4} , to 10^{-4} to 10^1 , preferably 10^{-3} to 10^1 , and most preferably 10^{-2} to 10^0 , may be achieved.

In the present invention, recombinase refers to a family of RecA-like and Rad51-like recombination proteins all having essentially all or most of the same functions, particularly: (i) the recombinase protein's ability to properly bind to and position targeting polynucleotides on their homologous targets and (ii) the ability of recombinase protein/targeting polynucleotide complexes to efficiently find and bind to complementary endogenous sequences. The best characterized RecA protein is from E coli, in addition to the wild-type protein a number of mutant RecA proteins have been identified (e.g., RecA803; see Madiraju et al., PNAS USA 85(18):6592 (1988); Madiraju et al, Biochem. 31:10529 (1992); Lavery et al., J. Biol. Chem. 267:20648 (1992)). Further, many organisms have RecA-like recombinases with strand-transfer activities (e.g., Fugisawa et al., (1985) Nucl. Acids Res. 13: 7473; Hsieh et al., (1986) Cell 44: 885; Hsieh et al., (1989) J. Biol. Chem. 264: 5089; Fishel et al., (1988) Proc. Natl. Acad. Sci. (USA) 85: 3683; Cassuto et al., (1987) Mol. Gen. Genet. 208: 10; Ganea et al., (1987) Mol. Cell Biol. 7: 3124; Moore et al., (1990) J. Biol. Chem. 19:11108; Keene et al., (1984) Nucl. Acids Res. 12: 3057; Kimeic, (1984) Cold Spring Harbor Symp. 48: 675; Kmeic, (1986) Cell 44:

545; Kolodner et al., (1987) *Proc. Natl. Acad. Sci. USA* 84: 5560; Sugino et al., (1985) *Proc. Natl. Acad. Sci. USA* 85: 3683; Halbrook et al., (1989) *J. Biol. Chem.* 264: 21403; Eisen et al., (1988) *Proc. Natl. Acad. Sci. USA* 85: 7481; McCarthy et al., (1988) *Proc. Natl. Acad. Sci. USA* 85: 5854; Lowenhaupt et al., (1989) *J. Biol. Chem.* 264: 20568, which are incorporated herein by reference. Examples of such recombinase proteins include, for example but not limited to: RecA, RecA803, uvsX, and other RecA mutants and RecA-like recombinases (Roca, A. I. (1990) *Crit. Rev. Biochem. Molec. Biol.* 25: 415), *sep1* (Kolodner et al. (1987) *Proc. Natl. Acad. Sci. (U.S.A.)* 84:5560; Tishkoff et al. *Molec. Cell. Biol.* 11:2593), *RuvC* (Dunderdale et al. (1991) *Nature* 354: 506), *DST2*, *KEM1*, *XRN 1* (Dykstra et al. (1991) *Molec. Cell. Biol.* 11:2583), *STPalph/DST1* (Clark et al. (1991) *Molec. Cell. Biol.* 11:2576), *HPP-1* (Moore et al. (1991) *Proc. Natl. Acad. Sci. (U.S.A.)* 88:9067), other target recombinases (Bishop et al. (1992) *Cell* 69 439; Shinohara et al. (1992) *Cell* 69 457); incorporated herein by reference. RecA may be purified from *E. coli* strains, such as *E. coli* strains JC 12772 and JC1 5369 (available from A.J. Clark and M. Madiraju, University of California-Berkeley, or purchased commercially). These strains contain the RecA coding sequences on a "runaway" replicating plasmid vector present at a high copy numbers per cell. The RecA803 protein is a high-activity mutant of wild-type RecA. The art teaches several examples of recombinase proteins, for example, from *Drosophila*, yeast, plant, human, and non-human mammalian cells, including proteins with biological properties similar to RecA (i.e., RecA-like recombinases), such as *Rad51*, *Rad55*, *Rad57*, *dmcl* from mammals and yeast. In addition, the recombinase may actually be a complex of proteins, i.e. a "recombinosome". In addition, included within the definition of a recombinase are portions or fragments of recombinases which retain recombinase biological activity, as well as variants or mutants of wild-type recombinases which retain biological activity, such as the *E. coli* RecA803 mutant with enhanced recombinase activity.

5.3.6.1. RecA or rad51

In a preferred embodiment, RecA or *rad51* is used. For example, RecA protein is typically obtained from bacterial strains that overproduce the protein: wild-type *E. coli* RecA protein and mutant RecA803 protein may be purified from such strains. Alternatively, RecA protein can also be purchased from, for example, Pharmacia (Piscataway, NJ) or Boehringer Mannheim (Indianapolis, Indiana).

RecA proteins, and its homologs, form a nucleoprotein filament when it coats a single-stranded DNA. In this nucleoprotein filament, one monomer of RecA protein is bound to about 3 nucleotides. This property of RecA to coat single-stranded DNA is essentially sequence independent, although particular sequences favor initial loading of RecA onto a polynucleotide (e.g., nucleation sequences). The nucleoprotein filament(s) can be formed on essentially any DNA molecule and can be formed in cells (e.g., mammalian cells), forming complexes with both single- stranded and double-stranded DNA, although the loading conditions for dsDNA are somewhat different than for ssDNA.

5.3.6.1.1. The Recombinase Is Combined With Targeting Polynucleotides

The recombinase is combined with targeting polynucleotides as is more fully outlined below. By "nucleic acid" or "oligonucleotide" or "polynucleotide" or grammatical equivalents herein means at least two nucleotides covalently linked together. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramidate (Beaucage et al., *Tetrahedron* 49(10):1925 (1993) and references therein; Letsinger, *J. Org. Chem.* 35:3800 (1970); Sprinzl et al., *Eur. J. Biochem.* 81:579 (1977); Letsinger et al., *Nucl. Acids Res.* 14:3487 (1986); Sawai et al, *Chem. Lett.* 805 (1984), Letsinger et al., *J. Am. Chem. Soc.* 110:4470 (1988); and Pauwels et al., *Chemica Scripta* 26:141 91986)), phosphorothioate, phosphorodithioate, O-methylphosphoroamidite linkages (see Eckstein, *Oligonucleotides and Analogues: A Practical Approach*, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, *J. Am. Chem. Soc.* 114:1895 (1992); Meier et al., *Chem. Int. Ed. Engl.* 31:1008 (1992); Nielsen, *Nature*, 365:566 (1993); Carlsson et al., *Nature* 380:207 (1996), all of which are incorporated by reference). These modifications of the ribose-phosphate backbone or bases may be done to facilitate the addition of other moieties such as chemical constituents, including 2' O-methyl and 5' modified substituents, as discussed below, or to increase the stability and half-life of such molecules in physiological environments.

The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, both genomic and cDNA, RNA or a hybrid, where the nucleic acid contains any

combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthine and hypoxanthine, etc. Thus, for example, chimeric DNA-RNA molecules may be used such as described in Cole-Strauss et al., *Science* 273:1386 (1996) and Yoon et al., *PNAS USA* 93:2071 (1996), both of which are hereby incorporated by reference.

In general, the targeting polynucleotides may comprise any number of structures, as long as the changes do not substantially effect the functional ability of the targeting polynucleotide to result in homologous recombination. For example, recombinase coating of alternate structures should still be able to occur.

By "targeting polynucleotides" herein is meant the polynucleotides used to make alterations in the protein domains as described herein. Targeting polynucleotides are generally ssDNA or dsDNA, most preferably two complementary single-stranded DNAs.

Targeting polynucleotides are generally at least about 5 to 2000 nucleotides long, preferably about 12 to 200 nucleotides long, at least about 200 to 500 nucleotides long, more preferably at least about 500 to 2000 nucleotides long, or longer; however, as the length of a targeting polynucleotide increases beyond about 20,000 to 50,000 to 400,000 nucleotides, the efficiency of transferring an intact targeting polynucleotide into the cell decreases. The length of homology may be selected at the discretion of the practitioner on the basis of the sequence composition and complexity of the predetermined endogenous target DNA sequence(s) and guidance provided in the art, which generally indicates that 1.3 to 6.8 kilobase segments of homology are preferred when non-recombinase mediated methods are utilized (Hasty et al. (1991) *Molec. Cell. Biol.* 11: 5586; Shulman et al. (1990) *Molec. Cell. Biol.* 10: 4466, which are incorporated herein by reference).

Targeting polynucleotides have at least one sequence that substantially corresponds to, or is substantially complementary to, a predetermined endogenous DNA sequence. As used herein, the terms "predetermined target nucleic acid" and "predetermined target sequence" and "predetermined domain of a target nucleic acid" refer to polynucleotide sequences contained in a target nucleic acid. Such sequences include, for example, chromosomal sequences (e.g., structural genes, regulatory sequences including promoters and enhancers,

recombinatorial hotspots, repeat sequences, integrated proviral sequences, hairpins, palindromes), episomal or extrachromosomal sequences (e.g., replicable plasmids or viral replication intermediates) including chloroplast and mitochondrial DNA sequences. By "predetermined" or "pre-selected" it is meant that the target sequence may be selected at the discretion of the practitioner on the basis of known or predicted sequence information, and is not constrained to specific sites recognized by certain site-specific recombinases (e.g., FLIP recombinase or CRE recombinase). In some embodiments, the predetermined endogenous DNA target sequence will be other than a naturally occurring germline DNA sequence (e.g., a transgene, parasitic, mycoplasmal or viral sequence). An exogenous polynucleotide is a polynucleotide which is transferred into a target cell but which has not been replicated in that host cell; for example, a virus genome polynucleotide that enters a cell by fusion of a virion to the cell is an exogenous polynucleotide, however, replicated copies of the viral polynucleotide subsequently made in the infected cell are endogenous sequences (and may, for example, become integrated into a cell chromosome). Similarly, transgenes which are microinjected or transfected into a cell are exogenous polynucleotides, however integrated and replicated copies of the transgene(s) are endogenous sequences.

5.3.6.1.1.1. Target Nucleic Acid Comprises A Nucleotide Sequence Encoding A Protein Or Polypeptide Or Can Be Made To Comprise Non-Coding Regions As Well

In a preferred embodiment, the target nucleic acid comprises a nucleotide sequence encoding a protein or polypeptide, although as outlined herein, target nucleic acids may be made to non-coding regions as well. By "protein" herein is meant at least two covalently attached amino acids, which includes proteins, polypeptides, oligopeptides and peptides. Thus "amino acid" or "peptide residue", as used herein means naturally occurring and naturally modified amino acids. For example, "amino acid" also includes imino acid residues such as proline and hydroxyproline. A "naturally modified amino acid" includes for examples, amino acids that are modified to contain carbohydrate structures, such as high-mannose or complex carbohydrates, phosphate, or lipids. In the preferred embodiment, the amino acids are in the (S) or L-configuration.

The nucleotide sequence encoding the polypeptide is preferably operably linked to transcription and translation control elements operable in a host cell of interest, such that,

introduction of the target nucleic acid results in expression of the encoded protein. The transcription control elements include a promoter, such as, a constitutive or inducible promoter. When the host cell of interest is a eukaryotic cell, enhancer elements are optionally employed. In a preferred embodiment the target nucleic acid is an extrachromosomal vector such as a plasmid. In other embodiments, the target nucleic acid is a viral vector, such as, a retrovirus, a phage, a BAC, PAC, YAC, MAC or other types of genomic and chromosomal DNA.

The term "naturally-occurring" as used herein as applied to an object refers to the fact that an object can be found in nature. For example, a polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring.

5.3.6.1.1.2. The Target Nucleic Acid Comprises A Nucleic Acid Encoding A Protein Domain

The methods of the invention are used for alteration and evolution of protein domains; that is, in a preferred embodiment, the target nucleic acid comprises a nucleic acid encoding a protein domain. By "protein domain" and grammatical equivalents as used herein are meant a region of a protein that provides a specific structural and/or functional characteristic. Accordingly, a protein domain is an enzymatic active site, a ligand binding site, an allosteric effector region, an epitope, a region of a protein that is modified, such as, by addition of a carbohydrate, phosphate or lipid. A domain also relates to the hydrophobicity or hydrophilicity of a region and, therefore, also includes extracellular, intracellular, and transmembrane domains. Cell targeting sequences, such as, a signal peptide, nuclear localization sequence, mitochondrial localization sequences, etc. that direct proteins to either an extracellular or subcellular locale are domains. Additional domains include regions of proteins that interact with other proteins or nucleic acids, for example, include multimerization sequences, zinc-finger motifs, and the like. In another aspect, a protein domain is a region encoded by an exon.

Targeting polynucleotides have at least one sequence that substantially corresponds to, or is substantially complementary to, a target nucleic acid; in a preferred embodiment, it corresponds or complements a nucleic acid encoding a protein domain. By "corresponds

to" herein is meant that a polynucleotide sequence is homologous (i.e., may be similar or identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the complementary sequence can hybridize to all or a portion of a reference polynucleotide sequence. Thus, one of the complementary single stranded targeting polynucleotides is complementary to one strand of the endogenous target domain sequence (i.e. Watson) and corresponds to the other strand of the endogenous target domain sequence (i.e. Crick). Thus, the complementarity between two single-stranded targeting polynucleotides need not be perfect. For illustration, the nucleotide sequence "TATAC" corresponds to a reference sequence "TATAC" and is perfectly complementary to a reference sequence "GTATA."

The terms "substantially corresponds to" or "substantial identity" or "homologous" as used herein denotes a characteristic of a nucleic acid sequence, wherein a nucleic acid sequence has at least about 50 percent sequence identity as compared to a reference sequence, typically at least about 70 percent sequence identity, and preferably at least about 85 percent sequence identity as compared to a reference sequence. The percentage of sequence identity is calculated excluding small deletions or additions which total less than 25 percent of the reference sequence. The reference sequence may be a subset of a larger sequence, such as a portion of a gene or flanking sequence, or a repetitive portion of a chromosome. However, the reference sequence is at least 18 nucleotides long, typically at least about 30 nucleotides long, and preferably at least about 50 to 100 nucleotides long.

"Substantially complementary" as used herein refers to a sequence that is complementary to a sequence that substantially corresponds to a reference sequence. In general, targeting efficiency increases with the length of the targeting polynucleotide portion that is substantially complementary to a reference sequence present in the target DNA.

By "sequence homology" herein is meant sequence similarity or sequence identity.

Nucleic acid similarity can be determined using, for example, BLASTN (Altschul et al. 1990. J. Mol. Biol. 147:195-197). BLASTN uses a simple scoring system in which matches count +5 and mismatches -4. To achieve computational efficiency, the default parameters have been incorporated directly into the source code.

5.3.7. Percent Nucleic Acid Sequence Identity Is Determined

In an alternative embodiment, percent nucleic acid sequence identity is determined. In percent identity calculations relative weight is not assigned to the various types of sequence variation, such as, insertions, deletions, substitutions, etc. Only identities are scored positively (+1) and all forms of sequence variation given a value of "0", which obviates the need for a weighted scale or parameters as described above for sequence similarity calculations. Percent sequence identity can be calculated, for example, by dividing the number of matching identical residues by the total number of residues of the "shorter" sequence in the aligned region and multiplying by 100. The "longer" sequence is the one having the most actual residues in the aligned region.

5.3.8. Domain Homology Clamps: A Portion Of The Targeting Polynucleotide That Can Specifically Hybridize To A Nucleic Acid Encoding A Domain Within A Gene Of Interest

These corresponding/complementary sequences are sometimes referred to herein as "domain homology clamps", as they serve as templates for homologous pairing with the predetermined endogenous sequence(s). Thus, a "domain homology clamp" is a portion of the targeting polynucleotide that can specifically hybridize to a nucleic acid encoding a domain within a gene of interest. "Specific hybridization" is defined herein as the formation of hybrids between a targeting polynucleotide (e.g., a polynucleotide of the invention which may include substitutions, deletion, and/or additions as compared to the predetermined target nucleic acid sequence) and a predetermined target nucleic acid, wherein the targeting polynucleotide preferentially hybridizes to the predetermined target nucleic acid such that, for example, at least one discrete band can be identified on a Southern blot of nucleic acid prepared from target cells that contain the target nucleic acid sequence, and/or a targeting polynucleotide in an intact nucleus localizes to a discrete chromosomal location characteristic of a unique or repetitive sequence. As will be appreciated by those in the art, a target domain sequence may be present in more than one target polynucleotide species (e.g., a particular target sequence may occur in multiple members of a gene family). It is evident that optimal hybridization conditions will vary depending upon the sequence composition and length(s) of the targeting polynucleotide(s) and target(s), and the experimental method selected by the practitioner. Various guidelines

may be used to select appropriate hybridization conditions (see, Maniatis et al., *Molecular Cloning: A Laboratory Manual* (1989), 2nd Ed., Cold Spring Harbor, N.Y. and Berger and Kimmel, *Methods in Enzymology*. Volume 152, *Guide to Molecular Cloning Techniques* (1987), Academic Press, Inc., San Diego, CA.), which are incorporated herein by reference. Methods for hybridizing a targeting polynucleotide to a discrete chromosomal location in intact nuclei are known in the art, see for example WO 93/05177 and Kowalczykowski and Zarling (1994) in *Gene Targeting*, Ed. Manuel Vega.

In targeting polynucleotides, domain homology clamps are typically located at or near the 5' or 3' end, preferably domain homology clamps are internal or located at each end of the polynucleotide (Berinstein et al. (1992) *Molec. Cell. Biol.* 12: 360, which is incorporated herein by reference). Without wishing to be bound by any particular theory, it is believed that the addition of recombinases permits efficient gene targeting with targeting polynucleotides having short (i. e., about 10 to 1000 basepair long) segments of homology, as well as with targeting polynucleotides having longer segments of homology.

5.3.9. Targeting Polynucleotides

5.3.9.1. Targeting Polynucleotides That Have Domain Homology Clamps That Are Highly Homologous To The Predetermined Target Endogenous Domain Functional Domain Nucleic Acid Sequence

Therefore, it is preferred that targeting polynucleotides of the invention have domain homology clamps that are highly homologous to the predetermined target endogenous domain functional domain nucleic acid sequence(s). Typically, targeting polynucleotides of the invention have at least one domain homology clamp that is at least about 18 to 35 nucleotides long, and it is preferable that domain homology clamps are at least about 20 to 100 nucleotides long, and more preferably at least about 100-500 nucleotides long, although the degree of sequence homology between the domain homology clamp and the targeted sequence and the base composition of the targeted sequence will determine the optimal and minimal clamp lengths (e.g., G-C rich sequences are typically more thermodynamically stable and will generally require shorter clamp length). Therefore, both domain homology clamp length and the degree of sequence homology can only be determined with reference to a particular predetermined sequence, but domain homology clamps generally must be at least about 10 nucleotides long and must also substantially

correspond or be substantially complementary to a predetermined target sequence. Preferably, a homology clamp is at least about 10, and preferably at least about 50 nucleotides long and is substantially identical to or complementary to a predetermined target sequence. Without wishing to be bound by a particular theory, it is believed that the addition of recombinases to a targeting polynucleotide enhances the efficiency of homologous recombination between homologous, nonisogenic sequences (e.g., between an exon 2 sequence of an albumin gene of a Balb/c mouse and a homologous albumin gene exon 2 sequence of a C57/BL6 mouse), as well as between isogenic sequences.

5.3.9.2. Targeting Polynucleotides Comprising A Plurality Of Targeting Polynucleotides Comprising At Least One Shared Homology Clamp And A Degenerate Sequence

In one aspect of the invention, the targeting polynucleotides comprise a plurality of targeting polynucleotides comprising at least one shared homology clamp and a degenerate sequence. By "plurality" herein is meant more than one. The targeting polynucleotides find use in the mutagenesis and evolution of a target nucleic acid sequence that encodes specific protein domain by insertion, deletion and/or substitution of the nucleic acid sequence encoding the domain. In one embodiment the degenerate sequence is completely randomized, representing all possible combinations of nucleotides. In another embodiment, the degenerate sequence is biased, for example, to eliminate sequences encoding for transcriptional or translational stop signals. In another embodiment, the degenerate sequence is biased, to represent the codon bias of a host cell or class of organisms. The degenerate sequence is optionally biased to randomize specific sequence while maintaining other sequences constant. The length of the degenerate sequence is determined by the practitioner and is based on the desired number of nucleotides within the predetermined sequence to be modified.

5.3.9.3. Targeting Polynucleotides Substantially Identical To The Predetermined Target Sequence

In an alternative embodiment, the targeting polynucleotides are substantially identical to the predetermined target sequence. In the presence of a recombinase, the targeting polynucleotides form complexes with a predetermined target sequence of a target nucleic acid. As a part of the complex, the predetermined target sequence is resistant to nuclease

digestion. The regions flanking the polynucleotide:target complex are susceptible to single-strand specific exonucleases. Accordingly, to effect domain specific evolution, these regions are nicked and the resultant fragments are reassembled and recombined by PCR as described below and by Stemmer et al. Nature. 370:389-391 and Stemmer et al. PNAS USA 91:10747-10751, hereby incorporated by reference.

The formation of heteroduplex joints is not a stringent process; genetic evidence supports the view that the classical phenomena of meiotic gene conversion and aberrant meiotic segregation results in part from the inclusion of mismatched base pairs in heteroduplex joints, and the subsequent correction of some of these mismatched base pairs before replication. Observations on RecA protein have provided information on parameters that affect the discrimination of relatedness from perfect or near-perfect homology and that affect the inclusion of mismatched base pairs in heteroduplex joints. The ability of RecA protein to drive strand exchange past all single base-pair mismatches and to form extensively mismatched joints in superhelical DNA reflect its role in recombination and gene conversion. This error-prone process may also be related to its role in mutagenesis. RecA-mediated pairing reactions involving DNA of X 174 and G4, which are about 70 percent homologous, have yielded homologous recombinants (Cunningham et al. (1981) Cell 24: 213), although RecA preferentially forms homologous joints between highly homologous sequences, and is implicated as mediating a homology search process between an invading DNA strand and a recipient DNA strand, producing relatively stable heteroduplexes at regions of high homology. Accordingly, it is the fact that recombinases can drive the homologous recombination reaction between strands which are significantly, but not perfectly, homologous, which allows gene conversion and the modification of target sequences. Thus, targeting polynucleotides may be used to introduce nucleotide substitutions, insertions and deletions into an endogenous functional domain nucleic acid sequence, and thus the corresponding amino acid substitutions, insertions and deletions in proteins expressed from the endogenous domain functional domain nucleic acid sequence. By "endogenous" in this context herein is meant the naturally occurring sequence, i.e. sequences or substances originating from within a cell or organism. Similarly, "exogenous" refers to sequences or substances originating outside the cell or organism.

5.3.9.4. Method Where Two Substantially Complementary Targeting Polynucleotides Are Used

In a preferred embodiment, two substantially complementary targeting polynucleotides are used.

5.3.9.5. Method Where The Targeting Polynucleotides Form A Double Stranded Hybrid, Which May Be Coated With Recombinase

In one embodiment, the targeting polynucleotides form a double stranded hybrid, which may be coated with recombinase, although when the recombinase is RecA, the loading conditions may be somewhat different from those used for single stranded nucleic acids.

5.3.9.6. Method Where Two Substantially Complementary Single- Stranded Targeting Polynucleotides Are Used

In a preferred embodiment, two substantially complementary single- stranded targeting polynucleotides are used. The two complementary single-stranded targeting polynucleotides are usually of equal length, although this is not required. However, as noted below, the stability of the four strand hybrids of the invention is putatively related, in part, to the lack of significant unhybridized single-stranded nucleic acid, and thus significant unpaired sequences are not preferred. Furthermore, as noted above, the complementarity between the two targeting polynucleotides need not be perfect. The two complementary single-stranded targeting polynucleotides are simultaneously or contemporaneously introduced into a target cell harboring a predetermined endogenous target sequence, generally with at least one recombinase protein (e.g., RecA). Under most circumstances, it is preferred that the targeting polynucleotides are incubated with RecA or other recombinase prior to introduction into a target cell, so that the recombinase protein(s) may be "loaded" onto the targeting polynucleotide(s), to coat the nucleic acid, as is described below. Incubation conditions for such recombinase loading are described *infra*. A targeting polynucleotide may contain a sequence that enhances the loading process of a recombinase, for example a RecA loading sequence is the recombinogenic nucleation sequence poly[d(A-C)], and its complement, poly[d(G-T)]. The duplex sequence poly[d(A-C)*d(G-T)_n], where n is from 5 to 25, is a middle repetitive element in target DNA.

There appears to be a fundamental difference in the stability of RecA- protein-mediated D-loops formed between one single-stranded DNA (ssDNA) probe hybridized to negatively supercoiled DNA targets in comparison to relaxed or linear duplex DNA targets. Internally located dsDNA target sequences on relaxed linear DNA targets hybridized by ssDNA probes produce single D- loops, which are unstable after removal of RecA protein (Adzuma, *Genes Devel.* 6:1679 (1992); Hsieh et al, *PNAS USA* 89:6492 (1992); Chiu et al., *Biochemistry* 32:13146 (1993)). This probe DNA instability of hybrids formed with linear duplex DNA targets is most probably due to the incoming ssDNA probe W-C base pairing with the complementary DNA strand of the duplex target and disrupting the base pairing in the other DNA strand. The required high free-energy of maintaining a disrupted DNA strand in an unpaired ssDNA conformation in a protein-free single-D-loop apparently can only be compensated for either by the stored free energy inherent in negatively supercoiled DNA targets or by base pairing initiated at the distal ends of the joint DNA molecule, allowing the exchanged strands to freely intertwine. However, the addition of a second complementary ssDNA to the three- strand-containing single-D-loop stabilizes the deproteinized hybrid joint molecules by allowing W-C base pairing of the probe with the displaced target DNA strand. The addition of a second RecA-coated complementary ssDNA (cssDNA) strand to the three-strand containing single D-loop stabilizes deproteinized hybrid joints located away from the free ends of the duplex target DNA (Sena & Zarling, *Nature Genetics* 3:365 (1993); Revet et al. *J. Mol. Biol.* 232:779 (1993); Jayasena and Johnston, *J. Mol. Bio.* 230:1015 (1993)). The resulting four-stranded structure, named a double D-loop by analogy with the three-stranded single D-loop hybrid has been shown to be stable in the absence of RecA protein. This stability likely occurs because the restoration of W-C basepairing in the parental duplex would require disruption of two W-C basepairs in the double-D-loop (one W-C pair in each heteroduplex D-loop).

Since each base-pairing in the reverse transition (double-D-loop to duplex) is less favorable by the energy of one W-C basepair, the pair of cssDNA probes are thus kinetically trapped in duplex DNA targets in stable hybrid structures. The stability of the double-D loop joint molecule within internally located probe:target hybrids is an intermediate stage prior to the progression of the homologous recombination reaction to

the strand exchange phase. The double D-loop permits isolation of stable multistranded DNA recombination intermediates.

In addition, when the targeting polynucleotides are used to generate insertions or deletions in an endogenous nucleic acid sequence, as is described herein, the use of two complementary single-stranded targeting polynucleotides allows the use of internal homology clamps. The use of internal homology clamps allows the formation of stable deproteinized cssDNA:probe target hybrids with homologous DNA sequences containing either relatively small or large insertions and deletions within a homologous DNA target. Without being bound by theory, it appears that these probe:target hybrids, with heterologous inserts in the cssDNA probe, are stabilized by the re-annealing of cssDNA probes to each other within the double-D-loop hybrid, forming a novel DNA structure with an internal homology clamp. Similarly stable double-D-loop hybrids formed at internal sites with heterologous inserts in the linear DNA targets (with respect to the cssDNA probe) are equally stable. Because cssDNA probes are kinetically trapped within the duplex target, the multi-stranded DNA intermediates of homologous DNA pairing are stabilized and strand exchange is facilitated.

5.3.10. Length Of The Internal Homology Clamp (i. e. The Length Of The Insertion Or Deletion)

In a preferred embodiment, the length of the internal homology clamp (i. e. the length of the insertion or deletion) is from about 1 to 50% of the total length of the targeting polynucleotide, with from about 1 to about 20% being preferred and from about 1 to about 10% being especially preferred, although in some cases the length of the deletion or insertion may be significantly larger. As for the domain homology clamps, the complementarity within the internal homology clamp need not be perfect.

A targeting polynucleotide used in a method of the invention typically is a single-stranded nucleic acid, usually a DNA strand, or derived by denaturation of a duplex DNA, which is complementary to one (or both) strand(s) of the target duplex nucleic acid. Thus, one of the complementary single stranded targeting polynucleotides is complementary to one strand of the endogenous target sequence (i.e. Watson) and the other complementary single stranded targeting polynucleotide is complementary to the other strand of the endogenous target sequence (i.e. Crick). The domain homology clamp sequence preferably contains at

least 90-95% sequence homology with the target sequence (although as outlined above, less sequence homology can be tolerated), to insure sequence-specific targeting of the targeting polynucleotide to the endogenous DNA domain target. Each single-stranded targeting polynucleotide is typically about 50-600 bases long, although a shorter or longer polynucleotide may also be employed.

5.3.11. Method For Making The Targeting Polynucleotides

Once the gene family and domain sequence is selected, the targeting polynucleotides are made, as will be appreciated by those in the art. For example, for large targeting polynucleotides, plasmids are engineered to contain an appropriately sized gene sequence with a deletion or insertion in the gene of interest and at least one flanking homology clamp which substantially corresponds or is substantially complementary to an endogenous target DNA sequence. Vectors containing a targeting polynucleotide sequence are typically grown in *E. coli* and then isolated using standard molecular biology methods. Alternatively, targeting polynucleotides may be prepared in single-stranded form by oligonucleotide synthesis methods, which may first require, especially with larger targeting polynucleotides, formation of subfragments of the targeting polynucleotide, typically followed by splicing of the subfragments together, typically by enzymatic ligation. In general, as will be appreciated by those in the art, targeting polynucleotides may be produced by chemical synthesis of oligonucleotides, nick-translation of a double-stranded DNA template, polymerase chain-reaction amplification of a sequence (or ligase chain reaction amplification), purification of prokaryotic or target cloning vectors harboring a sequence of interest (e.g., a cloned cDNA or genomic clone, or portion thereof) such as plasmids, phagemids, YACs, cosmids, bacteriophage DNA, other viral DNA or replication intermediates, or purified restriction fragments thereof, as well as other sources of single and double-stranded polynucleotides having a desired nucleotide sequence. When using microinjection procedures it may be preferable to use a transfection technique with linearized sequences containing only modified target gene sequence and without vector or selectable sequences. The modified gene site is such that a homologous recombinant between the exogenous targeting polynucleotide and the endogenous DNA target sequence can be identified by using carefully chosen primers and PCR, followed by analysis to detect if PCR products specific to the desired targeted event are present (Erlich et al., (1991) *Science* 252: 1643, which is incorporated herein by reference). Several

studies have already used PCR to successfully identify and then clone the desired transfected cell lines (Zimmer and Gruss, (1989) Nature 338: 150; Mouellic et al., (1990) Proc. Natl. Acad. Sci. USA 87: 4712; Shesely et al., (1991) Proc. Natl. Acad. Sci. USA 88: 4294, which are incorporated herein by reference). This approach is very effective when the number of cells receiving exogenous targeting polynucleotide(s) is high (i.e., with microinjection, or with liposomes) and the treated cell populations are allowed to expand to cell groups of approximately 1×10^4 cells (Capecchi, (1989) Science 244: 1288). When the target gene is not on a sex chromosome, or the cells are derived from a female, both alleles of a gene can be targeted by sequential inactivation (Mortensen et al., (1991) Proc. Natl. Acad. Sci. US 88: 7036). Alternatively, animals heterologous for the target gene can be bred to homologously as is known in the art.

The invention may also be practiced with individual targeting polynucleotides which do not comprise part of a complementary pair. In each case, a targeting polynucleotide is introduced into a target cell simultaneously or contemporaneously with a recombinase protein, typically in the form of a recombinase coated targeting polynucleotide as outlined herein (i.e., a polynucleotide pre-incubated with recombinase wherein the recombinase is noncovalently bound to the polynucleotide; generally referred to in the art as a nucleoprotein filament).

5.3.12. Alterations In The Target Nucleic Acid Comprising A Domain Or Domains Of Interest

The present invention allows for the introduction of alterations in the target nucleic acid comprising a domain or domains of interest. That is, the fact that heterologies are tolerated in targeting polynucleotides allows for two things: first, the use of a heterologous domain homology clamps that may target genes encoding functional domains of a protein or multiple proteins, resulting in a variety of genotypes and phenotypes, and secondly, the introduction of alterations to the target sequence. Thus typically, a targeting polynucleotide (or complementary polynucleotide pair) has a portion or region having a sequence that is not present in the preselected endogenous targeted sequence(s) (i.e., a nonhomologous portion or mismatch) which may be as small as a single mismatched nucleotide, several mismatches, or may span up to about several kilobases or more of nonhomologous sequence.

5.3.12.1. Methods And Compositions For Inactivation Of A Domain Of A Gene

Accordingly, in a preferred embodiment, the methods and compositions of the invention are used for inactivation of a domain of a gene. That is, exogenous targeting polynucleotides can be used to inactivate, decrease or alter the biological activity of one or more domains in a gene of a cell (or transgenic nonhuman animal or plant). This finds particular use in the generation of animal models of disease states, or in the elucidation of gene function and activity, similar to "knock out" experiments. Alternatively, the biological activity of the wild-type gene may be either decreased, or the wild-type activity altered to mimic disease states. This includes genetic manipulation of non-coding gene sequences that affect the transcription of genes, including, promoters, repressors, enhancers and transcriptional activating sequences.

5.3.12.1.1. Amino Acid Substitutions, Insertions Or Deletions In The Endogenous Target Sequences

Thus in a preferred embodiment, homologous recombination of the targeting polynucleotide and endogenous target sequence will result in amino acid substitutions, insertions or deletions in the endogenous target sequences, potentially both within the functional domain region and outside of it, for example as a result of the incorporation of PCR tags. This will generally result in modulated or altered gene function of the endogenous gene, including both a decrease or elimination of function as well as an enhancement of function. Nonhomologous portions are used to make insertions, deletions, and/or replacements in a predetermined endogenous targeted DNA sequence, and/or to make single or multiple nucleotide substitutions in a predetermined endogenous target DNA sequence so that the resultant recombined sequence (i.e., a targeted recombinant endogenous sequence) incorporates some or all of the sequence information of the nonhomologous portion of the targeting polynucleotide(s). Thus, the nonhomologous regions are used to make variant sequences, i.e. targeted sequence modifications. In this way, site directed modifications may be done in a variety of systems for a variety of purposes.

5.3.12.1.1.1. Disruption By Either The Substitution, Insertion, Deletion Or Frame Shifting Of Nucleotides

The endogenous target sequence, generally nucleic acid encoding a domain, may be disrupted in a variety of ways. The term "disrupt" as used herein comprises a change in the coding or non-coding sequence of an endogenous nucleic acid. In one preferred embodiment, a disrupted gene will no longer produce a functional gene product. In another preferred embodiment, a disrupted gene produces a variant gene product. Generally, disruption may occur by either the substitution, insertion, deletion or frame shifting of nucleotides.

5.3.12.1.1.2. Disruption By Amino Acid Substitutions

In one embodiment, amino acid substitutions are made. This can be the result of either the incorporation of a non-naturally occurring domain sequence into a target, or of more specific changes to a particular sequence outside of the domain sequence.

5.3.12.1.1.3. Disruption By An Insertion Sequence

In one embodiment, the endogenous sequence is disrupted by an insertion sequence. The term "insertion sequence" as used herein means one or more nucleotides which are inserted into an endogenous gene to disrupt it. In general, insertion sequences can be as short as 1 nucleotide or as long as a gene, as outlined herein. For non-gene insertion sequences, the sequences are at least 1 nucleotide, with from about 1 to about 50 nucleotides being preferred, and from about 10 to 25 nucleotides being particularly preferred. An insertion sequence may comprise a polylinker sequence, with from about 1 to about 50 nucleotides being preferred, and from about 10 to 25 nucleotides being particularly preferred. Insertion sequence may be a PCR tag used for identification of the first gene. In a preferred embodiment, an insertion sequence comprises a gene which not only disrupts the endogenous gene, thus preventing its expression, but also can result in the expression of a new gene product. Thus, in a preferred embodiment, the disruption of an endogenous gene by an insertion sequence gene is done in such a manner to allow the transcription and translation of the insertion gene. An insertion sequence that encodes a gene may range from about 50 bp to 5000 bp of cDNA or about 5000 bp to 50000 bp of genomic DNA. As will be appreciated by those in the art, this can be done in a variety of ways. In a preferred embodiment, the insertion gene is targeted to the endogenous gene in

such a manner as to utilize endogenous regulatory sequences, including promoters, enhancers or a regulatory sequence. In an alternate embodiment, the insertion sequence gene includes its own regulatory sequences, such as a promoter, enhancer or other regulatory sequence etc.

Particularly preferred insertion sequence genes include, but are not limited to, genes which encode selection or reporter proteins. In addition, the insertion sequence genes may be modified or variant genes.

5.3.12.1.1.4. Disruption By Deletions

The term "deletion" as used herein comprises removal of a portion of the nucleic acid sequence of an endogenous gene. Deletions range from about 1 to about 100 nucleotides, with from about 1 to 50 nucleotides being preferred and from about 1 to about 25 nucleotides being particularly preferred, although in some cases deletions may be much larger, and may effectively comprise the removal of the entire functional domain, the entire endogenous gene and/or its regulatory sequences. Deletions may occur in combination with substitutions or modifications to arrive at a final modified endogenous gene.

5.3.12.1.1.5. Disruption Simultaneously by An Insertion And A Deletion

In a preferred embodiment, endogenous genes may be disrupted simultaneously by an insertion and a deletion. For example, a domain of an endogenous gene, with or without its regulatory sequences, may be removed and replaced with an insertion sequence gene. Thus, for example, all but the regulatory sequences of an endogenous gene may be removed, and replaced with an insertion sequence gene, which is now under the control of the endogenous gene's regulatory elements.

The term "regulatory element" is used herein to describe a non-coding sequence which affects the transcription or translation of a gene including, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, enhancer or activator sequences, dimerizing sequences, etc. In a preferred embodiment, the regulatory sequences include a promoter and transcriptional start and stop sequence. Promoter sequences encode either constitutive or inducible

promoters. The promoters may be either naturally occurring promoters or hybrid promoters. Hybrid promoters, which combine elements of more than one promoter, are also known in the art, and are useful in the present invention.

In addition to domain homology clamps and optional internal homology clamps, the targeting polynucleotides of the invention may comprise additional components, such as cell-uptake components, chemical substituents, purification tags, etc.

5.3.12.2. Targeting Polynucleotide Comprising A Cell-Uptake Component

In a preferred embodiment, at least one of the targeting polynucleotides comprises at least one cell- uptake component. As used herein, the term "cell-uptake component" refers to an agent which, when bound, either directly or indirectly, to a targeting polynucleotide, enhances the intracellular uptake of the targeting polynucleotide into at least one cell type (e.g., hepatocytes). A targeting polynucleotide of the invention may optionally be conjugated, typically by covalently or preferably noncovalent , binding, to a cell-uptake component. Various methods have been described in the art for targeting DNA to specific cell types. A targeting polynucleotide of the invention can be conjugated to essentially any of several cell-uptake components known in the art. For targeting to hepatocytes, a targeting polynucleotide can be conjugated to an asialoorosomucoid (ASOR)-poly-L-lysine conjugate by methods described in the art and incorporated herein by reference (Wu GY and Wu CH (1987) J. Biol. Chem. 262:4429; Wu GY and Wu CH (1988) Biochemistry 27:887; Wu GY and Wu CH (1988) J. Biol. Chem. 263: 1462 1; Wu GY and Wu CH (1992) J. Biol. Chem. 267: 12436; Wu et al. (1991) J. Biol. Chem. 266: 14338; and Wilson et al. 0 992) J..Biol. Chem. 267: 963, W092/06180; W092/05250; and W091/17761, which are incorporated herein by reference).

Alternatively, a cell-uptake component may be formed by incubating the targeting polynucleotide with at least one lipid species and at least one protein species to form protein-lipid-polynucleotide complexes consisting essentially of the targeting polynucleotide and the lipid-protein cell-uptake component. Lipid vesicles made according to Feigner (W091/17424, incorporated herein by reference) and/or cationic lipidization (WO91/16024, incorporated herein by reference) or other forms for polynucleotide administration (EP 465,529, incorporated herein by reference) may also be

employed as cell-uptake components. Nucleases, DNA damaging chemicals, UV radiation or gamma- radiation may also be used.

In addition to cell-uptake components, targeting components such as nuclear localization signals may be used, as is known in the art. See for example Kido et al., *Exper. Cell Res.* 198:107-114 (1992), hereby expressly incorporated by reference.

Typically, a targeting polynucleotide of the invention is coated with at least one recombinase and is conjugated to a cell-uptake component, and the resulting cell targeting complex is contacted with a target cell under uptake conditions (e.g., physiological conditions) so that the targeting polynucleotide and the recombinase(s) are internalized in the target cell. A targeting polynucleotide may be contacted simultaneously or sequentially with a cell-uptake component and also with a recombinase; preferably the targeting polynucleotide is contacted first with a recombinase, or with a mixture comprising both a cell-uptake component and a recombinase under conditions whereby, on average, at least about one molecule of recombinase is noncovalently attached per targeting polynucleotide molecule and at least about one cell-uptake component also is noncovalently attached. Most preferably, coating of both recombinase and cell-uptake component saturates essentially all of the available binding sites on the targeting polynucleotide. A targeting polynucleotide may be preferentially coated with a cell-uptake component so that the resultant targeting complex comprises, on a molar basis, more cell-uptake component than recombinase(s). Alternatively, a targeting polynucleotide may be preferentially coated with recombinase(s) so that the resultant targeting complex comprises, on a molar basis, more recombinase(s) than cell- uptake component.

Cell-uptake components are included with recombinase-coated targeting polynucleotides of the invention to enhance the uptake of the recombinase-coated targeting polynucleotide(s) into cells, particularly for in vivo gene targeting applications, such as gene therapy to treat genetic diseases, including neoplasia, and targeted homologous recombination to treat viral infections wherein a viral sequence (e.g., an integrated hepatitis B virus (HBV) genome or genome fragment) may be targeted by homologous sequence targeting and inactivated. Alternatively, a targeting polynucleotide may be coated with the cell-uptake component and targeted to cells with a contemporaneous or

simultaneous administration of a recombinase (e.g., liposomes or immunoliposomes containing a recombinase, a viral-based vector encoding and expressing a recombinase).

In addition to recombinase and cellular uptake components, at least one of the targeting polynucleotides may include chemical substituents. Exogenous targeting polynucleotides that have been modified with appended chemical substituents may be introduced along with recombinase (e.g., RecA) into a metabolically active target cell to homologously pair with a predetermined endogenous DNA target sequence in the cell. In a preferred embodiment, the exogenous targeting polynucleotides are derivatized, and additional chemical substituents are attached, either during or after polynucleotide synthesis, respectively, and are thus localized to a specific endogenous target sequence where they produce an alteration or chemical modification to a local DNA sequence. Preferred attached chemical substituents include, but are not limited to: cross-linking agents (see Podymnugin et al., *Biochem.* 34:13098 (1995) and 35:7267 (1996), both of which are hereby incorporated by reference), nucleic acid cleavage agents, metal chelates (e.g., iron/EDTA chelate for iron catalyzed cleavage), topoisomerases, endonucleases, exonucleases, ligases, phosphodiesterases, photodynamic porphyrins, chemotherapeutic drugs (e.g., adriamycin, doxorubicin), intercalating agents, labels, base-modification agents, agents which normally bind to nucleic acids such as labels, etc. (see for example Afonina et al., *PNAS USA* 93:3199 (1996), incorporated herein by reference) immunoglobulin chains, and oligonucleotides. Iron/EDTA chelates are particularly preferred chemical substituents where local cleavage of a DNA sequence is desired (Hertzberg et al. (1982) *J. Am. Chem. Soc.* 104: 313; Hertzberg and Dervan (1984) *Biochemistry* 23: 3934; Taylor et al. (1984) *Tetrahedron* 40: 457; Dervan, PB (1986) *Science* 232: 464, which are incorporated herein by reference). Further preferred are groups that prevent hybridization of the complementary single stranded nucleic acids to each other but not to unmodified nucleic acids; see for example Kutryavin et al., *Biochem.* 35:11170 (1996) and Woo et al., *Nucleic Acid. Res.* 24(13):2470 (1996), both of which are incorporated by reference. 2'-O methyl groups are also preferred; see Cole-Strauss et al., *Science* 273:1386 (1996); Yoon et al., *PNAS* 93:2071 (1996)). Additional preferred chemical substituents include labeling moieties, including fluorescent labels. Preferred attachment chemistries include: direct linkage, e.g., via an appended reactive amino group (Corey and Schultz (1988) *Science* 238:1401, which is incorporated herein by reference)

and other direct linkage chemistries, although streptavidin/biotin and digoxigenin/antidigoxigenin antibody linkage methods may also be used. Methods for linking chemical substituents are provided in U.S. Patents 5,135,720, 5,093,245, and 5,055,556, which are incorporated herein by reference. Other linkage chemistries may be used at the discretion of the practitioner.

5.3.12.3. Targeting Polynucleotides Comprises At Least One Purification Tag Or Capture Moiety

In a preferred embodiment, at least one of the targeting polynucleotides comprises at least one purification tag or capture moiety, some of which are discussed above as chemical substituents, for example biotin, digoxigenin, psoralen, etc. Alternatively, the domain oligonucleotide could be directly attached to beads with the targeting reaction performed on a solid phase support.

5.3.12.4. Targeting Polynucleotides Are Coated With Recombinase Prior To Introduction To The Domain Target

In a preferred embodiment, the targeting polynucleotides are coated with recombinase prior to introduction to the domain target. The procedures below are directed to the use of *E. coli* RecA, although as will be appreciated by those in the art, other recombinases may be used as well. Targeting polynucleotides can be coated using GTPgammaS, mixes of ATPgammaS with rATP, rGTP and/or dATP, or dATP or rATP alone in the presence of an rATP generating system (Boehringer Mannheim). Various mixtures of GTPgammaS, ATPgammaS, ATP, ADP, dATP and/or rATP or other nucleosides may be used, particularly preferred are mixes of ATPgammaS and ATP or ATPgammaS and ADP. The targeting polynucleotide, whether double-stranded or single-stranded, is denatured by heating in an aqueous solution at 95- 100°C for five minutes, then placed in an ice bath for 20 seconds to about one minute followed by centrifugation at 0°C for approximately 20 sec, before use. When denatured targeting polynucleotides are not placed in a freezer at -20°C they are usually immediately added to standard RecA coating reaction buffer containing ATPgammaS, at room temperature, and to this is added the RecA protein.

Alternatively, RecA protein may be included with the buffer components and ATPgammaS before the polynucleotides are added.

RecA coating of targeting polynucleotide(s) is initiated by incubating polynucleotide-RecA mixtures at 37°C for 10-15 min. RecA protein concentration tested during reaction with polynucleotide varies depending upon polynucleotide size and the amount of added polynucleotide, and the ratio of RecA molecule: nucleotide preferably ranges between about 3:1 and 1:3. When single-stranded polynucleotides are RecA coated independently of their homologous polynucleotide strands, the mM and microM concentrations of ATPgammaS and RecA, respectively, can be reduced to one-half those used with double-stranded targeting polynucleotides (i.e., RecA and ATPgammaS concentration ratios are usually kept constant at a specific concentration of individual polynucleotide strand, depending on whether a single- or double-stranded polynucleotide is used).

RecA protein coating of targeting polynucleotides is normally carried out in a standard 10x RecA coating reaction buffer. 10x RecA reaction buffer (i.e., 10x AC buffer) consists of: 100 mM Tris acetate (pH 7.5 at 37°C), 20 mM magnesium acetate, 500 mM sodium acetate, 10 mM DTT, and 50% glycerol). All of the targeting polynucleotides, whether double-stranded or single-stranded, typically are denatured before use by heating to 95-100°C for five minutes, placed on ice for one minute, and subjected to centrifugation (10,000 rpm) at 0°C for approximately 20 seconds (e.g., in a Tomy centrifuge). Denatured targeting polynucleotides usually are added immediately to room temperature RecA coating reaction buffer mixed with ATPgammaS and diluted with double-distilled H₂O as necessary.

A reaction mixture typically contains the following components: (i) 0.2- 4.8 mM ATPgammaS; and (ii) between 1-100 ng/ul of targeting polynucleotide. To this mixture is added about 1-20, ul of RecA protein per 10-100 ul of reaction mixture, usually at about 2-10 mg/ml (purchased from Pharmacia or purified), and is rapidly added and mixed. The final reaction volume-for RecA coating of targeting polynucleotide is usually in the range of about 10-500 ul. RecA coating of targeting polynucleotide is usually initiated by incubating targeting polynucleotide-RecA mixtures at 37°C for about 10-15 min. RecA protein concentrations in coating reactions varies depending upon targeting polynucleotide size and the amount of added targeting polynucleotide: RecA protein concentrations are typically in the range of 5 to 50 uM. When single-stranded targeting polynucleotides are

coated with RecA, independently of their complementary strands, the concentrations of ATPgammaS and RecA protein may optionally be reduced to about one-half of the concentrations used with double-stranded targeting polynucleotides of the same length: that is, the RecA protein and ATPgammaS concentration ratios are generally kept constant for a given concentration of individual polynucleotide strands.

5.3.12.4.1. Evaluation Of Coating Of Targeting Polynucleotides With RecA Protein

The coating of targeting polynucleotides with RecA protein can be evaluated in a number of ways. First, protein binding to DNA can be examined using band-shift gel assays (McEntee et al., (1981) 1. Biol. Chem. 256: 8835). Labeled polynucleotides can be coated with RecA protein in the presence of ATPgammaS and the products of the coating reactions may be separated by agarose gel electrophoresis.

Following incubation of RecA protein with denatured duplex DNAs the RecA protein effectively coats single-stranded targeting polynucleotides derived from denaturing a duplex DNA. As the ratio of RecA protein monomers to nucleotides in the targeting polynucleotide increases from 0, 1:27, 1:2.7 to 3.7:1 for 121-mer and 0, 1:22, 1:2.2 to 4.5:1 for 159-mer, targeting polynucleotide's electrophoretic mobility decreases, i.e., is retarded, due to RecA-binding to the targeting polynucleotide. Retardation of the coated polynucleotide's mobility reflects the saturation of targeting polynucleotide with RecA protein. An excess of RecA monomers to DNA nucleotides is required for efficient RecA coating of short targeting polynucleotides (Leahy et al., (1986) J. Biol. Chem. 261: 954).

A second method for evaluating protein binding to DNA is in the use of nitrocellulose fiber binding assays (Leahy et al., (1986) J. Biol. Chem. 261:6954; Woodbury, et al., (1983) Biochemistry 22(20):4730-4737. The nitrocellulose filter binding method is particularly useful in determining the dissociation-rates for protein:DNA complexes using labeled DNA. In the filter binding assay, DNA:protein complexes are retained on a filter while free DNA passes through the filter. This assay method is more quantitative for dissociation-rate determinations because the separation of DNA:protein complexes from free targeting polynucleotide is very rapid.

Alternatively, recombinase protein(s) (prokaryotic, eukaryotic or endogenous to the target cell) may be exogenously induced or administered to a target cell simultaneously or contemporaneously (i.e., within about a few hours) with the targeting polynucleotide(s). Such administration is typically done by micro-injection, although electroporation, lipofection, and other transfection methods known in the art may also be used.

Alternatively, recombinase-proteins may be produced in vivo. For example, they may be produced from a homologous or heterologous expression cassette in a transfected cell or targeted cell, such as a transgenic totipotent cell (e.g. a fertilized zygote) or an embryonal stem cell (e.g., a murine ES cell such as AB-1) used to generate a transgenic non-human animal line or a somatic cell or a pluripotent hematopoietic stem cell for reconstituting all or part of a particular stem cell population (e.g. hematopoietic) of an individual.

Conveniently, a heterologous expression cassette includes a modulatable promoter, such as an ecdysone-inducible promoter- enhancer combination, an estrogen-induced promoter-enhancer combination, a CMV promoter- enhancer, an insulin gene promoter, or other cell-type specific, developmental stage-specific, hormone-inducible drug inducible, or other modulatable promoter construct so that expression of at least one species of recombinase protein from the cassette can be modulated for transiently producing recombinase(s) in vivo simultaneous or contemporaneous with introduction of a targeting polynucleotide into the cell. When a hormone-inducible promoter-enhancer combination is used, the cell must have the required hormone receptor present, either naturally or as a consequence of expression a co-transfected expression vector encoding such receptor. Alternatively, the recombinase may be endogenous and produced in high levels. In this embodiment, preferably in eukaryotic target cells such as tumor cells, the target cells produce an elevated level of recombinase. In other embodiments the level of recombinase may be induced by DNA damaging agents, such as mitomycin C, UV or gamma-irradiation. Alternatively, recombinase, levels may be elevated by transfection of a plasmid encoding the recombinase gene into the cell.

5.3.13. Specialized Applications

5.3.13.1. Identification of New Members Of Gene Families Which May Be Useful In Functional Genomic Studies As Well As In The Identification Of New Drug Targets

Once made, the compositions of the invention find use in a number of applications upon administration to target cells. In general, the compositions and methods of the invention

are useful to identify new members of gene families which may be useful in functional genomic studies as well as in the identification of new drug targets; both of these may be accomplished through the generation of "knock out" animal models. In addition, the present invention allows the modification of functional domain targets, the creation of transgenic plants and animals, the cloning of genes containing domain functional domains, etc.

5.3.13.2. Domain Specific Gene Evolution

Once made and administered to a target host cell, the compositions of the invention find use in a number of applications, including domain specific gene evolution. The polypeptide or protein encoded by the targeted nucleic undergoes homologous recombination with the plurality of polynucleotides to produce a plurality of modified target nucleic acids that are expressed to produce a plurality of modified proteins. Selection systems are employed to identify and isolate host cells expressing proteins having a desired property or phenotype. For example, if the expressed protein is an enzyme, cells having a modified enzyme activity are identified. The desired activity can be an increased or decreased or altered activity. Proteins having the desired phenotype are selected and isolated, the modified nucleic acid is sequenced to identify sequences effecting the desired activity, and the process is repeated iteratively as needed to produce a protein having a desired activity or property. In this and other embodiments, suitable target sequences include nucleic acid sequences encoding therapeutically or commercially relevant proteins, including, but not limited to, enzymes (proteases, recombinases, lipases, kinases, carbohydrases, isomerases, tautomerases, nucleases etc.), hormones, receptors, transcription factors, growth factors, cytokines, globin genes, immunosuppressive genes, tumor suppressors, oncogenes, complement- activating genes, milk proteins (casein, alpha-lactalbumin, beta-lactoglobulin, bovine and human serum albumin), immunoglobulins, milk proteins, and pharmaceutical proteins and vaccines.

In a preferred embodiment, the methods of the invention are used to generate pools or libraries of variant nucleic acid sequences, and cellular libraries containing the variant sequences. This idea is somewhat similar to the "gene shuffling" techniques of the literature (see Stemmer et al., 1994, Nature 370:389 which attempt to rapidly "evolve" genes by making multiple random changes simultaneously. In the present invention, this

end is accomplished by using at least one cycle, and preferably reiterative cycles, of enhanced homologous recombination with targeting polynucleotides containing random mismatches, substitutions, insertions, or deletions. By using a library of targeting polynucleotides comprising a plurality of random mutations, and repeating the homologous recombination steps as many times as needed, a rapid "gene evolution" can occur, wherein the new genes may contain large numbers of mutations.

Thus, in this embodiment, a plurality of targeting polynucleotides are used. The targeting polynucleotides each have at least one homology clamp that substantially corresponds to or is substantially complementary to the target sequence. Generally, the targeting polynucleotides are generated in pairs; that is, pairs of two single stranded targeting polynucleotides that are substantially complementary to each other are made (i.e. a Watson strand and a Crick strand). However, as will be appreciated by those in the art, less than a one to one ratio of Watson to Crick strands may be used; for example, an excess of one of the single stranded target polynucleotides (i.e. Watson) may be used. Preferably, sufficient numbers of each of Watson and Crick strands are used to allow the majority of the targeting polynucleotides to form double D-loops, which are preferred over single D-loops as outlined above. In addition, the pairs need not have perfect complementarity; for example, an excess of one of the single stranded target polynucleotides (i.e. Watson), which may or may not contain mismatches, may be paired to a large number of variant Crick strands, etc. Due to the random nature of the pairing, one or both of any particular pair of single- stranded targeting polynucleotides may not contain any mismatches. However, generally, at least one of the strands will contain at least one mismatch.

The plurality of pairs preferably comprise a pool or library of mismatches. The size of the library will depend on a number of factors, including the number of residues to be mutagenized, the susceptibility of the protein to mutation, etc., as will be appreciated by those in the art. Generally, a library in this instance preferably comprises at least 10% different mismatches over the length of the targeting polynucleotides, with at least 30% mismatches being preferred and at least 40% being particularly preferred, although as will be appreciated by those in the art, lower (1, 2, 5%, etc.) or higher amounts of mismatches being both possible and desirable in some instances. That is, the plurality of pairs comprise a pool of random and preferably degenerate mismatches over some regions or all

of the entire targeting sequence. As outlined herein, "mismatches" include substitutions, insertions and deletions, with the former being preferred. Thus, for example, a pool of degenerate variant targeting polynucleotides covering some, or preferably all, possible mismatches over some region are generated, as outlined above, using techniques well known in the art. Preferably, but not required, the variant targeting polynucleotides each comprise only one or a few mismatches (less than 10), to allow complete multiple randomization. That is, by repeating the homologous recombination steps any number of times, as is more fully outlined below, the mismatches from a plurality of probes can be incorporated into a single target sequence.

The mismatches can be either non-random (i.e. targeted) or random, including biased randomness. That is, in some instances specific changes are desirable, and thus the sequence of the targeting polynucleotides are specifically chosen. In a preferred embodiment, the mismatches are random. The targeting polynucleotides can be chemically synthesized, and thus may incorporate any nucleotide at any position. The synthetic process can be designed to generate randomized nucleic acids, to allow the formation of all or most of the possible combinations over the length of the nucleic acid, thus forming a library of randomized targeting polynucleotides. Preferred methods maximize library size and diversity.

It is important to understand that in any library system encoded by oligonucleotide synthesis one cannot have complete control over the codons that will eventually be incorporated into the peptide structure. This is especially true in the case of codons encoding stop signals (TAA, TGA, TAG). In a synthesis with NNN as the random region, there is a 3/64, or 4.69%, chance that the codon will be a stop codon. To alleviate this, random residues are encoded as NNK, where K= T or G. This allows for encoding of all potential amino acids (changing their relative representation slightly), but importantly preventing the encoding of two stop residues TAA and TGA.

5.3.13.2.1. Mismatches Are Fully Randomized, With No Sequence Preferences Or Constants At Any Position

In one embodiment, the mismatches are fully randomized, with no sequence preferences or constants at any position.

5.3.13.2.2. Biased Library

In a preferred embodiment, the library is biased. That is, some positions within the sequence are either held constant, or are selected from a limited number of possibilities. For example, in a preferred embodiment, the nucleotides or amino acid residues are randomized within a defined class, for example, of hydrophobic amino acids, hydrophilic residues, sterically biased (either small or large) residues, towards the creation of cysteines, for cross-linking, prolines for SH-3 domains, serines, threonines, tyrosines or histidines for phosphorylation sites, etc., or to purines, etc.

As will be appreciated by those in the art, the introduction of a pool of variant targeting polynucleotides (in combination with recombinase) to a target sequence, in vitro to an extrachromosomal sequence, can result in a large number of homologous recombination reactions occurring over time. That is, any number of homologous recombination reactions can occur on a single target sequence, to generate a wide variety of single and multiple mismatches within a single target sequence, and a library of such variant target sequences, most of which will contain mismatches and be different from other members of the library. This thus works to generate a library of mismatches.

5.3.13.2.2.1. Generating A Large Number Of Different Variants Within A Particular Region Of A Sequence, Similar To Cassette Mutagenesis But Not Limited By Sequence Length

In a preferred embodiment, the variant targeting polynucleotides are made to a particular region or domain of a sequence (i.e. a nucleotide sequence that encodes a particular protein domain). For example, it may be desirable to generate a library of all possible variants of a binding domain of a protein, without affecting a different biologically functional domain, etc. Thus, the methods of the present invention find particular use in generating a large number of different variants within a particular region of a sequence, similar to cassette mutagenesis but not limited by sequence length. This idea is sometimes referred to herein as "domain specific gene evolution". In addition, two or more regions may also be altered simultaneously using these techniques; thus "single domain" and "multi- domain" shuffling can be done. Suitable domains include, but are not limited to, kinase domains, nucleotide-binding sites, DNA binding sites, signaling domains, receptor

binding domains, transcriptional activating regions, promoters, origins, leader sequences, terminators, localization signal domains, and, in immunoglobulin genes, the complementarity determining regions (CDR), V_H , and V_L .

In a preferred embodiment, the variant targeting polynucleotides are made to the entire target sequence. In this way, a large number of single and multiple mismatches may be made in an entire sequence.

Thus, this embodiment proceeds as follows. A pool of , targeting polynucleotides are made each containing one or more mismatches. The probes are coated with recombinase as generally described herein, and introduced to the target sequence. Upon binding of the probes to form D-loops, the recombinase is preferably removed. These polynucleotide:target sequences can then introduced into recombinant proficient cells, to produce target protein which can then be tested for biological activity, based on the identification of the target sequence. Depending on the results, the altered target sequence can be used as the starting target sequence in reiterative rounds of homologous recombination, generally using the same library. Preferred embodiments utilize at least two rounds of homologous recombination, with at least 5 rounds being preferred and at least 10 rounds being particularly preferred. Again, the number of reiterative rounds that are performed will depend on the desired end-point, the resistance or susceptibility of the protein to mutation, the number of mismatches in each probe, etc.

5.3.14. Target Sequence

5.3.14.1. Target Sequences - An Immunoglobulin

In a preferred embodiment, the target sequence is an immunoglobulin. The amino terminal region of the light and heavy chains of an antibody that come together to form the antigen binding site and the variability of their amino acid sequences provides the structural basis for the diversity of antigen binding sites. The variability of the variable regions of both the heavy and light chains is for the most part restricted to three small hypervariable regions in each chain. The remaining part of the variable regions, known as framework regions, is relatively constant. Each of the hypervariable regions consists of only about 5 to 10 amino acids; the corresponding regions in the DNA encoding these regions are known as the complementarity determining regions, or CDRs. Thus to engineer an antibody library, for

example an antibody phage library, one can change the sequences in the CDR regions of both the heavy and light chains. Different permutations and combinations of CDRs can be changed and evolved to engineer antibody-phage libraries.

5.3.14.2. Target Sequence - A Single-Chain Fv Framework For Any Number Of Specific Antigens

In a preferred embodiment, the target sequence is a single-chain Fv framework for any number of specific antigens. Single chain Fv (scFv) consists of V_L and V_H domains of an immunoglobulin linked by a peptide spacer and thus contains the minimal antigen-binding domains of an antibody.

5.3.14.3. Target Sequence - An Antibody-Phage Fusion

In a preferred embodiment, antibody-phage fusions are used as the target sequence. As is known in the art, single-chain Fv fusions with the phage minor coat protein allows expression of the antibody on the surface of a phage, wherein it is available to bind antigen. Five copies of phage are expressed on the surface of the phage. It is therefore possible to express five scFv on the phage. This antibody-phage display system has been used previously to isolate novel antibodies. By starting with antibodies to any antigen, higher affinity antibodies may be made, as well as novel antibodies.

5.3.14.4. Target Sequence - The Coding Sequence For beta-Lactamase

In a preferred embodiment, the target sequence is the coding sequence for beta-lactamase.

Thus, the methods of the invention may be used to create superior recombinant reporter genes such as lacZ and green fluorescent protein (GFP); superior antibiotic and drug resistance genes; superior recombinase genes; superior recombinant vectors; and other superior recombinant genes and proteins, including immunoglobulins, vaccines or other proteins with therapeutic value. For example, targeting polynucleotides containing any number of alterations may be made to one or more functional or structural domains of a protein, and then the products of homologous recombination evaluated.

Once made and administered to target cells, the target cells may be screened to identify a cell that contains the targeted sequence modification. This will be done in any number of

ways, and will depend on the target gene and targeting polynucleotides as will be appreciated by those in the art. The screen may be based on phenotypic, biochemical, genotypic, or other functional changes, depending on the target sequence. In an additional embodiment, as will be appreciated by those in the art, selectable markers or marker sequences may be included in the targeting polynucleotides to facilitate later identification.

5.3.15. Kits Containing The Compositions Of The Invention Are Provided

In a preferred embodiment, kits containing the compositions of the invention are provided. The kits include the compositions, particularly those of libraries or pools of degenerate cDNA probes, along with any number of reagents or buffers, including recombinases, buffers, ATP, etc.

5.3.16. Targeting Polynucleotide:Target Nucleic Acid Complexes Serve As Substrates For Single-Stranded Endonucleases

In an alternate embodiment, the targeting polynucleotide:target nucleic acid complexes serve as substrates for single-stranded endonucleases, such as, S1 and mung bean nuclease. Preferably the targeting polynucleotides are substantially complementary and form double D-loops with the target nucleic acid. The junctions of the complexes are single-stranded in nature, and thus are susceptible to single-strand specific nucleases and junction-specific nucleases. Accordingly, treatment of the complex with a single-strand nuclease results in defined nicks in the selected region encoding a predetermined domain of a protein encoded by the target nucleic acid. The nicked target nucleic acid is disassociated from the targeting polynucleotides and are reassembled and "shuffled" in vitro by PCR (Stemmer. 1994. Nature 370:389-391) to produce a plurality modified nucleic acids. The modified nucleic acids are introduced into an appropriate host cell, as described above, for expression of the plurality of modified proteins. Selection techniques are used as described herein to identify and isolate a cell expressing a modified protein. The process is repeated iteratively as needed to further evolve the targeted nucleic acid.

5.3.17. Isolation Of New Members Of Gene Families That Comprise Particular Domains

In a preferred embodiment, the present invention finds use in the isolation of new members of gene families that comprise particular domains. The use of domain filaments (i.e. domain homology clamps preferably containing a purification tag such as biotin, disoxisenin, or one purification method such as the use of a RecA antibody), allows the identification of genes containing the domain. Once identified, the new genes can be cloned, sequenced and the protein gene products purified. As will be appreciated by those in the art, the functional importance of the new genes can be assessed in a number of ways, including functional studies on the protein level, as well as the generation of "knock out" animal models. By choosing domain sequences for therapeutically relevant protein domains, novel targets can be identified that can be used in screening of drug candidates.

5.3.18. Utilizing The Purification Tag To Isolate The Gene(s)

Thus, in a preferred embodiment, the present invention provides methods for isolating new members of gene families containing protein domains comprising introducing targeting polynucleotides comprising domain homology clamps and at least one purification tag, preferably biotin, to a mix of nucleic acid, such as a plasmid cDNA library or a cell, and then utilizing the purification tag to isolate the gene(s). The exact methods will depend on the purification tag; a preferred method utilizes the attachment of the binding ligand for the tag to a bead, which is then used to pull out the sequence. Alternatively anti-RecA antibodies could be used to capture RecA-coated probes. The genes are then cloned, sequenced, and reassembled if necessary, as is well known in the art.

5.3.19. Use In Functional Genomic Studies, By Providing The Creation Of Transgenic Animal Models Of Disease

In an alternate preferred embodiment, the present invention finds use in functional genomic studies, by providing the creation of transgenic animal models of disease. Thus, for example, domain sequences used in homologous recombination methods can generate animals that have a wide variety of mutations in a wide variety of related domains of genes, potentially resulting in a wide variety of phenotypes, including phenotypes related to disease states. That is, by targeting a domain family, one, two or multiple genes in the family may be altered in any given experiment, thus creating a wide variety of genotypes

and phenotypes to evaluate. Thus, in a preferred embodiment, the compositions and methods of the invention are used to generate pools or libraries of variant nucleic acid sequences, wherein the mutations are within the functional domain coding region, cellular libraries containing the variant libraries, and libraries of animals containing the variant libraries.

Furthermore, domain targeting can be used in cells or animals that are diseased or altered; in essence, domain targeting can be done to identify "reversion" genes, genes that can modulate disease states caused by domains of different genes. Thus for example the loss of one type of enzymatic activity, resulting in a disease phenotype, may be compensated by alterations in a different but homologous enzymatic activity.

Accordingly, once the recombinase-targeting polynucleotide compositions are formulated, they are introduced or administered into target cells. The administration is typically done as is known for the administration of nucleic acids into cells, and, as those skilled in the art will appreciate, the methods may depend on the choice of the target cell. Suitable methods include, but are not limited to, microinjection, electroporation, lipofection, etc. By "target cells" herein is meant prokaryotic or eukaryotic cells. Suitable prokaryotic cells include, but are not limited to, bacteria such as *E coli*, *Bacillus* species, and the extremophile bacteria such as thermophiles, halophiles, etc. Preferably, the prokaryotic target cells are recombination competent. Suitable eukaryotic cells include, but are not limited to, fungi such as yeast and filamentous fungi, including species of *Aspergillus*, *Trichoderma*, and *Neurospora*; plant cells including those of corn, sorghum, tobacco, canola, soybean, cotton, tomato, potato, alfalfa, sunflower, etc.; and animal cells, including fish, reptiles, amphibians, birds and mammals. Suitable fish cells include, but are not limited to, those from species of salmon, trout, tilapia, tuna, carp, flounder, halibut, swordfish, cod and zebrafish. Suitable bird cells include, but are not limited to, those of chickens, ducks, quail, pheasants, ostrich, and turkeys, and other jungle fowl or game birds. Suitable mammalian cells include, but are not limited to, cells from horses, cows, buffalo, deer, sheep, rabbits, rodents such as mice, rats, hamsters and guinea pigs, goats, pigs, primates, marine mammals including dolphins and whales, as well as cell lines, such as human cell lines of any tissue or stem cell type, and stem cells, including pluripotent and non-pluripotent, and non-human zygotes. Particular human cells including, but are not limited to, tumor cells of all types (particularly melanoma, myeloid leukemia, carcinomas of the

lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell), mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoietic, neural, skin, lung, kidney, liver and myocyte stem cells, osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to, Jurkat T cells, mouse La, HT1080, C127, Rat2, CV-1, NIH3T3 cells, CHO, COS, 293 cells, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

5.3.20. Procaryotic Cells Are Used To Identify, Clone, Or Alter Target Sequences

In a preferred embodiment, procaryotic cells are used to identify, clone, or alter target sequences, preferably protein domains. In this embodiment, a pre-selected target DNA sequence is chosen for alteration. Preferably, the pre-selected target DNA sequence is contained within an extrachromosomal sequence. By "extrachromosomal sequence" herein is meant a sequence separate from the chromosomal or genomic sequences. Preferred extrachromosomal sequences include plasmids (particularly procaryotic plasmids such as bacterial plasmids), pl vectors, viral genomes, yeast, bacterial and mammalian artificial chromosomes (YAC, BAC and MAC, respectively), and other autonomously self-replicating sequences, although this is not required. As described herein, a recombinase and at least two single stranded targeting polynucleotides which are substantially complementary to each other, each of which contain a homology clamp to the target sequence contained on the extrachromosomal sequence, are added to the extrachromosomal sequence, preferably in vitro. The two single stranded targeting polynucleotides are preferably coated with recombinase, and at least one of the targeting polynucleotides contain at least one nucleotide substitution, insertion or deletion. The targeting polynucleotides then bind to the target sequence in the extrachromosomal sequence to effect homologous recombination and form an altered extrachromosomal sequence which contains the substitution, insertion or deletion. The altered extrachromosomal sequence is then introduced into the procaryotic cell using techniques known in the art. Preferably, the recombinase is removed prior to introduction into the target cell, using techniques known in the art. For example, the reaction may be treated with proteases such as proteinase K, detergents such as SDS, and phenol extraction

(including phenol:chloroform:isoamyl alcohol extraction). These methods may also be used for eukaryotic cells. The cells are then grown under conditions which allow the expression of the variant nucleic acids to form variant proteins, particularly with alterations in domains.

5.3.20.1. Proteins Having The Desired Phenotype Are Selected And Isolated

In a preferred embodiment, proteins having the desired phenotype are selected and isolated, the modified nucleic acid is sequenced to identify sequences effecting the desired activity, and the process is repeated iteratively as needed to produce a protein having a desired activity or property. Thus, in a preferred embodiment, the methods of the invention are repeated until the desired protein or phenotype is seen.

Alternatively, the pre-selected target DNA sequence is a chromosomal sequence. In this embodiment, the recombinase with the targeting polynucleotides are introduced into the target cell, preferably eukaryotic target cells. In this embodiment, it may be desirable to bind (generally non-covalently) a nuclear localization signal to the targeting polynucleotides to facilitate localization of the complexes in the nucleus. See for example Kido et al., *Exper. Cell Res.* 198:107-114 (1992), hereby expressly incorporated by reference. The targeting polynucleotides and the recombinase function to effect homologous recombination, resulting in altered chromosomal or genomic sequences.

5.3.21. Eukaryotic Cells Are Used

In a preferred embodiment, eukaryotic cells are used. Basically, any mammalian cells may be used, with mouse, rat, primate and human cells being particularly preferred. Accordingly, suitable cell types include, but are not limited to, tumor cells of all types, i.e., fibroblasts, epithelial cells (particularly melanoma, myeloid leukemia, carcinomas of the lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell), mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoietic, neural, skin, lung, kidney, liver and myocyte stem cells (for use in screening for differentiation and de-differentiation factors), osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to,

Jurkat T cells, NIH 3T3 cells, CHO, Cos, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

For making transgenic non-human animals (which include homologously targeted non-human animals) embryonal stem cells (ES cells), donor cells for nuclear transfer and fertilized zygotes are preferred. In a preferred embodiment, embryonal stem cells are used. Murine ES cells, such as AB-1 line grown on mitotically inactive SNL76/7 cell feeder layers (McMahon and Bradley, Cell 62: 1073-1085 (1990)) essentially as described (Robertson, E.J. (1987) in *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach*. E.J. Robertson, ed. (Oxford: IRL Press), p. 71-112; Zijlstra et al., Nature 342:435-438 (1989); and Schwartzberg et al., Science 246:799-803 (1989), each of which is incorporated herein by reference) may be used for homologous gene targeting. Other suitable ES lines include, but are not limited to, the E14 line (Hooper et al. (1987) Nature 326: 292-295), the D3 line (Doetschman et al. (1985) J. Embryol. Exp. Morph. 87: 21-45), and the CCE line (Robertson et al. (1986) Nature 323: 445-448). The success of generating a mouse line from ES cells bearing a specific targeted mutation depends on the pluripotency of the ES cells (i.e., their ability, once injected into a host blastocyst, to participate in embryogenesis and contribute to the germ cells of the resulting animal).

The pluripotency of any given ES cell line can vary with time in culture and the care with which it has been handled. The only definitive assay for pluripotency is to determine whether the specific population of ES cells to be used for targeting can give rise to chimeras capable of germline transmission of the ES genome. For this reason, prior to gene targeting, a portion of the parental population of AB-1 cells is injected into C57B1/6J blastocysts to ascertain whether the cells are capable of generating chimeric mice with extensive ES cell contribution and whether the majority of these chimeras can transmit the ES genome to progeny.

5.3.22. Non-Human Zygotes Are Used

In a preferred embodiment, non-human zygotes are used, for example to make transgenic animals, using techniques known in the art (see U.S. Patent No. 4,873,191; Brinster et al., PNAS 86:7007 (1989); Susulic et al., J. Biol. Chem. 268:29483 (1993), and Cavard et al., Nucleic Acids Res. 16:2099 (1988), hereby incorporated by reference). Preferred zygotes

include, but are not limited to, animal zygotes, including fish, avian, reptilian, amphibian and mammalian zygotes. Suitable fish zygotes include, but are not limited to, those from species of salmon, trout, tuna, carp, flounder, halibut, swordfish, cod, tilapia and zebrafish. Suitable bird zygotes include, but are not limited to, those of chickens, ducks, quail, pheasant, turkeys, and other jungle fowl and game birds. Suitable mammalian zygotes include, but are not limited to, cells from horses, cows, buffalo, deer, sheep, rabbits, rodents such as mice, rats, hamsters and guinea pigs, goats, pigs, primates, and marine mammals including dolphins and whales. See Hogan et al., *Manipulating the Mouse Embryo (A Laboratory Manual)*, 2nd Ed. Cold Spring Harbor Press, 1994, incorporated by reference.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, micro-injection is commonly utilized for target cells, although calcium phosphate treatment, electroporation, lipofection, biolistics or viral-based transfection also may be used. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, and others (see, generally, Sambrook et al. *Molecular Cloning: A Laboratory Manual*, 2d ed., 1989, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., which is incorporated herein by reference). Direct injection of DNA and/or recombinase-coated targeting polynucleotides into target cells, such as skeletal or muscle cells also may be used (Wolff et al. (1990) *Science* 247: 1465, which is incorporated herein by reference).

5.3.23. Precursor Animals Or Cells Already Contain A Disease Allele

In a preferred embodiment, the precursor animals or cells already contain a disease allele. As used herein, the term "disease allele" refers to an allele of a gene which is capable of producing a recognizable disease. A disease allele may be dominant or recessive and may produce disease directly or when present in combination with a specific genetic background or pre-existing pathological condition. A disease allele may be present in the gene pool or may be generated de novo in an individual by somatic mutation. For example and not limitation, disease alleles include: activated oncogenes, a sickle cell anemia allele, a Tay-Sachs allele, a cystic fibrosis allele, a Lesch-Nyhan allele, a retinoblastoma-susceptibility allele, a Fabry's disease allele, a Huntington's chorea allele, and a xenoderma

pigmentosa allele. As used herein, a disease allele encompasses both alleles associated with human diseases and alleles associated with recognized veterinary diseases. For example, the deltaF508 CFTR allele in a human disease allele which is associated with cystic fibrosis in North Americans.

Once made and administered to target cells, new domains of genes may be isolated as outlined herein.

Alternatively, the target cells may be screened to identify a cell that contains the targeted functional domain sequence modification. This will be done in any number of ways, and will depend on the target domain and targeting polynucleotides as will be appreciated by those in the art. The screen may be based on phenotypic, biochemical, genotypic, or other functional changes, depending on the target sequence. For example, IgE levels may be evaluated for inflammation or asthma; vascular tone or blood pressure can be evaluated for hypertension, behavior screens can be done for neurologic effects, lipoprotein profiles can be screened for cardiovascular effects; secreted molecules can be evaluated for endocrine processes; CBCs can be done for hematology studies, etc. In an additional embodiment, as will be appreciated by those in the art, selectable markers or marker sequences may be included in the targeting polynucleotides to facilitate later identification.

The broad scope of this invention is best understood with reference to the following examples, which are not intended to limit the invention in any manner. All patents, patent applications, and publications cited herein are expressly incorporated by reference in their entirety.

5.4. Gene Deletion In Bacteria

This invention relates to a method and means for deleting a gene from a bacterial chromosome in a single step.

5.4.1 Applications

5.4.1.1 The Construction Of Special Bacterial Strains Which Have A Particular Genetic Background

Many applications require the construction of special bacterial strains which have a particular genetic background. These genetic backgrounds are the framework in which specific recombinant DNA plasmid constructions are tested to determine whether they can provide functions which are missing from the background of the bacteria. If such functions are provided by the recombinant plasmid, then there is positive evidence that a particular genetic locus or loci is encoded by the plasmid. The construction of genetic backgrounds is therefore a vital step in the subsequent cloning and investigation of specific genes.

5.4.1.2. The Gene In Question Has Been Deleted, So There Is No Production Whatsoever Of A Mutant Protein Making Analysis Much Less Ambiguous

The most common backgrounds are those in which a single mutation is present in a specific gene on the bacterial chromosome. This may result in synthesis of a defective version of the protein encoded by that gene, resulting in a specific cellular dysfunction. Correction of the cellular dysfunction, by introduction of a specific recombinant plasmid, is evidence that the relevant gene has been cloned onto the specific plasmid. However, a mutant protein may not be silent and may undergo interactions with other components, thereby creating the appearance that the plasmid gene encodes the entire active protein when it does not. This can seriously confuse the analysis. A much less ambiguous and therefore more desirable approach is one in which the gene in question has been deleted. In these circumstances, there is no production whatsoever of a mutant protein.

5.4.1.3. If Bacteria Is Being Used To Genetically Engineer A Protein, Deletion Of The Gene That Manufactures A Contaminating Protein Made By The Same Bacteria Can Reduce Purification Steps

Another situation where it is desirable to delete a complete gene from the chromosome of a bacteria is when the bacteria is being used in the production of a genetically engineered

protein. Examples of these situations include the expression of insulin, growth hormone, protein A, and various vaccines from recombinant genes inserted into *E. coli*. Many times the *E. coli* produces proteins which contaminate the purified product produced by the genetic engineering. Although it is possible to add additional purification steps to remove this contaminant, it would be preferable to avoid the problem entirely by deleting the gene encoding the contaminating protein. Methods that are presently used to alter a gene include random mutation or inactivation of the gene sequence by mutation, insertion, or deletion of some portion of the gene. However, this can still lead to the production of inactive protein fragments or deletion of more of the chromosome than is necessary or desirable.

It is therefore an object of the present invention to provide a method and means for deleting a specific gene from a bacteria.

It is another object of the present invention to provide a method and means for inserting and/or inactivating or deleting the *recA* gene in a variety of bacteria.

5.4.2. Miscellaneous Applications

5.4.2.1 Creating A Deletion In A Defined Target Within The Bacterial Chromosome

There are two obstacles which have to be overcome. One is to develop a method which will create a deletion in a defined target within the bacterial chromosome.

5.4.2.2 Generalizing The Approach So That Even Essential Genes Can Be Deleted

The second is to generalize the approach so that even essential genes can be deleted. This is because many of the genes of interest are essential ones. The deletion of an essential gene normally would result in cell death.

Essential proteins include enzymes of glycolysis, enzymes associated with amino acid or sugar biosynthesis, enzymes and factors associated with protein and nucleic acid biosynthesis (including both RNA and DNA), enzymes required for the synthesis of cofactors for oxidation, reduction, methylation and transamination processes, and enzymes necessary for synthesis of essential lipids and polysaccharides or of any other essential

molecule, including various nucleic acids, such as transfer or ribosomal RNAs, and segments of nucleic acids, such as gene regulatory elements.

It is therefore an object of the present invention to provide a method wherein an organism is produced which does not contain genetic material coding for the molecule which is to be cloned and expressed in the organism.

A further object of the present invention is to provide a method whereby a deficient organism which is to be used for cloning an essential gene remains viable even under restrictive conditions.

5.4.3 Specialized Applications

5.4.3.1. Method For The Deletion Of A Gene From A Bacteria Using A Single Step Procedure That Is Applicable To Any Gene That Has Been Cloned

Disclosed is a method for the deletion of a gene from a bacteria using a single step procedure that is applicable to any gene that has been cloned. The procedure depends upon site-directed recombination of linear DNA fragments with sequences on the chromosome as a function of *recA* in combination with the subsequent inactivation or deletion of the *recA* gene. The method is analogous to a procedure used to give insertions by homologous recombination into specific plasmid genes.

5.4.3.1.1. Strategy For Construction Of Chromosomal Deletions

The basic strategy for construction of chromosomal deletions is to transform the bacteria with linear DNA fragments which contain an antibiotic resistant or other phenotypically detectable gene segment (a "marker") flanked by sequences homologous to a closely spaced region on the cell chromosome containing the gene to be deleted. A double-crossover event within the homologous sequences, effectively deleting the entire gene, is selected for by screening for the antibiotic resistant phenotype.

The linear fragment is not integrated into the chromosome in the absence of enzymes expressed by the *recA* gene. Accordingly, if the gene is absent or inactive, a *recA* gene must be inserted into the cell prior to the linear recombination event, and then inactivated or removed to prevent subsequent incorporation of other non-chromosomal sequences into

the chromosome. This is particularly important if the bacteria is used as a host for the expression of genetically engineered proteins from sequences carried on plasmids or other extrachromosomal elements. The *recA* gene can be provided either in the form of an extrachromosomal element such as a plasmid or through incorporation of the gene into the chromosome. The *recA* gene is preferably inactivated or deleted by means of a double reciprocal recombination event utilizing linear sequences containing sequences homologous to the flanking sequences on either side of the *recA* gene in the chromosome. This is essentially the same method used to delete or insert a gene into the chromosome by homologous recombination as described above.

The present invention includes isolated linear DNA fragments constructed for use in the method for deleting a gene from the chromosome and for inserting or deleting the *recA* gene.

The method and sequences are applicable to a variety of bacteria including strains of *Escherichia*, *Pseudomonas*, *Agrobacterium*, *Proteus*, *Erwinia*, *Shigella*, *Bacillus*, *Rhizobium*, *Vibrio*, *Salmonella*, *Streptococcus*, and *Haemophilus*.

A plasmid which has a temperature-sensitive replicon and a wild-type allele of the desired gene is used to restore or maintain the phenotype produced by the deleted gene. This plasmid maintains production of the desired protein, and therefore cell viability if the encoded protein is essential to cell growth, when the chromosomal copy of the desired gene has been deleted. However, since the resulting cells have a temperature-sensitive phototype, the expression of the plasmid gene may be easily prevented by culturing the host strain at an elevated temperature. The resulting deficient host strain may then be used to screen other mutated and cloned genes for their ability to produce the desired protein.

5.5. Additional Considerations

5.5.1. Maintaining The Viability Of The Bacteria

The present invention is a method for deleting any gene from a bacterial strain, while maintaining the viability of the bacteria if the gene encodes an essential molecule and deletion of the essential gene results in a lethal phenotype. The gene to be deleted is provided on a plasmid with a temperature sensitive replicon. The cells now have a temperature sensitive phenotype. When the cells are grown at an elevated temperature under conditions which allow rapid detection of the absence of the desired molecule, complementation of this phenotype by introduction of a DNA fragment fused onto a stable plasmid is strong evidence for cloning of the gene which has been deleted from the chromosome.

5.5.2. Homologous Recombination

The present invention is a method for deleting any gene from a bacterial strain employing linear DNA fragments incorporating sequences homologous to the sequences flanking the gene to be deleted in the chromosome and sequences allowing insertion or removal/inactivation of *recA* in a variety of bacteria.

5.5.2.1. Roles For Several Enzymes Needed In Recombination

Homologous recombination has been detected in a wide variety of organisms, from simple bacteriophages to complex eukaryotic cells. Genetic and biochemical investigations have defined roles for several enzymes needed in recombination.

5.5.2.2. RecA

5.5.2.2.1. Alignment Of DNA Molecules Before Exchange

The RecA protein participates in the early steps of synapse, allowing alignment of DNA molecules before exchange, in strand transfer, where there is transfer of a single-stranded segment to a recipient duplex to form a limited heteroduplex region between the interacting DNAs and in the extension of this heteroduplex region by a reaction involving the concerted winding and unwinding of incoming and outgoing DNA chains, respectively. The hydrolysis of ATP by RecA protein is required for these events in vitro.

5.5.2.2.2. Controlling Expression Of A Group Of Unlinked Genes That Aid In Recovery Of Cells After Exposure To DNA Damaging Agents

The *recA* gene performs another equally important role in cell metabolism by controlling expression of a group of unlinked genes that aid in recovery of cells after exposure to DNA-damaging agents. This response, termed the SOS response, involves genes that participate in repair of DNA damage, mutagenesis, and coordination of cell division events.

5.5.2.2.3. Characterization

The purified RecA protein is a single polypeptide ranging in weight from about 37,000 to about 42,000. Although there is some variation in the sequences between bacterial strains, the *recA* proteins from a variety of bacteria in general have been isolated and characterized by interspecies complementation and assays utilizing comparisons with isolated, characterized proteins.

The present method and the linear fragments for use in the present method are not limited to organisms such as *E. coli*. The *recA* gene is found in a variety of bacteria including both gram negative and gram positive organisms. The *recA* gene has been isolated or identified and characterized in several organisms. It is possible to prepare a genomic library from any bacterial species and to isolate a clone containing a sequence homologous to a characterized *recA* gene by interspecific complementation using one of the available DNA clones containing a *recA* sequence or cross-reactivity with antisera to RecA proteins from a well-characterized organism such as *E. coli*.

RecA⁺ and RecA⁻ strains are available from a variety of sources. For example, cloning and characterization of *recA* genes and *recA* proteins from *Proteus vulgaris*, *Erwinia carotovora*, *Shigella flexneria* and *Escherichia coli* are described by S. L. Keener, et al., in *J. Bacter.* 160(1), 153-160 (1984). The RecA proteins produced by these organisms were demonstrated to be highly conserved among the species. In fact, the protein produced by one species could be introduced into another species where it complemented repair and regulatory defects of *recA* mutations. Other bacterial *recA* genes and gene products have been described by C.A. Miles, et al, in *Mol. Gen.Genet.* 204,161-165 (1986) (*Agrobacterium tumefaciens* C58), I. Goldberg et al, *J. Bacteriol.* 165(3), 715-722(1986),

(*Vibrio cholera*), M. Better et al, J. Bacteriol. 155(1), 311-316 (1983), (*Rhizobium meliloti*), T.A. Kokjohn et al, J. Bacteriol. 163(2), 568-572 (1985), (*Pseudomonas aeruginosa*), and C.M. Lovett, Jr. et al, J. Bio.Chem. 260(6), 3305-3313 (1985) (*Bacillus subtilis*). These articles detail the isolation and characterization of gene libraries and the proteins encoded by the *recA* genes using techniques known to those skilled in the art including construction of gene libraries, identification of homologous genes using hybridization to probes from other more well characterized species such as *E. coli*, isolation and characterization of RecA proteins using antisera to RecA proteins from *E. coli*, and interspecies complementation of deficient strains of *E. coli* using gene segments from the libraries. The isolated proteins were useful for in vitro complementation studies. RecA deficient strains and RecA clones are available from many of the laboratories cited in the above articles and from the *E. coli* Genetic Stock Center at Yale University run by Dr. Barbara Bachman.

5.5.2.3. Construction Of Linear DNA Fragments Which Have Sequences Homologous To Closely Spaced Regions On The Chromosome In The Bacteria Which Flank Each Side Of The Gene Of Interest

The method whereby the DNA fragment is introduced into the chromosome to delete a gene or to regulate *recA* involves the construction of linear DNA fragments which have sequences homologous to closely spaced regions on the chromosome in the bacteria which flank each side of the gene of interest. In general, they must contain at least 80 to 100 nucleotides homologous to sequences flanking the gene to be deleted or the *recA* gene. An antibiotic resistant locus such as *Kan^r*, which encodes a protein making the bacterial resistant to the antibiotic, or other marker, is placed between these flanking sequences.

The linear DNA segment is introduced into a specific cell strain and selection is done for a double, reciprocal recombination which will delete the target gene and insert the marker into its place. As a result of the insertion of the antibiotic resistant gene, the cells are now resistant to the antibiotic. Cells which do not contain the insertion are eliminated by growing the bacteria in a medium containing the antibiotic.

Any other gene which allows for rapid screening of the cells containing a double, reciprocal recombination may be used in place of a gene for antibiotic resistance. For example, any gene for an essential protein, including enzymes, cofactors, proteins which are necessary for the synthesis of essential lipids, polysaccharides, nucleic acids, and other protein molecules such as receptors, as well as nucleic acids which have functional activity such as ribozymes, may be used. Other genes which confer a detectable phenotype on the cell strain such as sensitivity to temperature or ultraviolet radiation, auxotrophism for a sugar, amino acid, protein or nucleotide, or any other phenotype which can be detected by chemical indicators either in vitro or in vivo assay, or an immunoassay for a specific cellular component may also be used. Such chemical, radioactive, or immunological screening assays are well known to those skilled in the art.

5.5.3. Eliminating Contamination From The Host Producing Its Own Protein By Deletion And Replacement

In one application, in which the goal is to produce and purify a foreign protein, and the microorganism encodes its own version of the protein, the gene for the microorganism's own protein is eliminated. The gene for the foreign protein is then inserted and the protein produced. The purification process is thereby simplified since there is no contamination by the host protein, whether analogous to the protein being produced or unrelated which copurifies with, or interferes with the purification of, the protein being produced. For example, a protein may interfere with binding of the protein to be purified to a column.

5.5.4. A Plasmid Is Used To Introduce A Mutated Gene Into An Organism

In a second application, a plasmid is used to introduce a gene into an organism which typically contains a mutation in the gene to be investigated resulting in a negative phenotype for the product of the gene to be investigated. Failure of the organism to produce a biologically active form of the protein encoded by the mutated gene may confer a lethal phenotype under certain defined conditions. For example, this can be at a temperature, designated as the restrictive temperature, at which the mutant protein denatures or otherwise undergoes inactivation. The cloned gene which is introduced is selected for by virtue of its ability to confer cell viability or any other detectable phenotype for the desired protein at the restrictive temperature. The acquisition of viability

or other detectable phenotype at the normally restrictive temperature is evidence that the gene of interest has been cloned.

5.5.4.1. Problems: The Defective Host Protein May Interact With A Protein Produced From The Introduced Plasmid

The major problem with this second system is that, unless the inactive gene is deleted in entirety, the defective host protein may interact with a protein produced from the introduced plasmid. This interaction may stabilize the defective host protein enough so that its activity is restored even at the restrictive temperature. In this case, the restoration of growth at the restrictive temperature would be a false positive, that is, the growth would not be due to activity encoded by the cloned DNA segment. This problem holds true for all selections based on complementation of a phenotype which is due to a defect in a specific protein. Such phenotypes include temperature sensitivity, amino acid auxotrophies or any other auxotrophies which result from a lack of synthesis of a key ingredient such as a sugar, nucleotide, critical protein or nucleic acid, or cofactor used for oxidation, reduction, or transamination reactions.

5.5.5. False Positives

The problem of "false positives" also exists for cloned DNA pieces which are created to encode enzyme fragments as a means to define the catalytic core or to define any segment which achieves a specific purpose, such as a piece which undergoes self-association, binds to a specific ligand or receptor, or forms a specific complex or array with one or more additional components. In these cases, the engineering of protein fragments, which are tested in a host cell that encodes a defective version of the protein of interest, is seriously hampered if the defective host protein interacts in any way with the engineered pieces.

5.5.6. Use Of Temperature Sensitive Replicons

In the present invention, specifically designed linear DNA fragments are used to create a deletion of a gene by site-specific recombination. These fragments are transformed into the host cell. Cell viability or the detectable phenotype can be maintained during the procedure by provision of the gene encoding the desired protein on a recombinant plasmid that has a temperature-sensitive replicon, so that the cells which contain the deletion have a temperature sensitive phenotype. To achieve the deletion by recombination with the

linear DNA fragments, it is necessary for the cells to have a RecA⁺ phenotype which is derived from recA, or its equivalent. Once recombination has occurred, the cell must immediately be changed to RecA⁻ or else the temperature sensitive plasmid will recombine with homologous sequences on the chromosome. The same would apply to any other extrachromosomal element where integration into the host chromosome would be undesirable.

The RecA⁻ phenotype may be achieved by simultaneous inactivation of recA during the transformation with linear fragments or, after the transformation, by immediately introducing RecA⁻ by mating with an appropriate RecA⁻ strain or by transduction with a phage which carries a RecA⁻ gene segment. Although mutagenesis may also be an effective means of making the cell RecA⁻, this is a "hit or miss" approach. The preferred method is to use homologous recombination of linear DNA sequences bounded by sequences hybridizing to the sequences flanking the recA gene. The recA gene is necessary in order for the gene encoding the desired protein to be incorporated into the organism. However, any plasmids or other extrachromosomal elements in the cell will be incorporated unless the recA gene is immediately removed. This is a particular concern where the bacteria serves as a host for the expression of a genetically engineered protein from multicopy plasmids.

6. Detection Methods (Screening/Selection)

6.1. Basic approach: artificially evolving cells to acquire a new or improved property by stochastic &/or non-stochastically mutagenizing

6.1.1. General: successive cycles of recombination and screening/selection

The invention provides methods for artificially evolving cells to acquire a new or improved property by recursive sequence recombination. Briefly, recursive sequence recombination entails successive cycles of recombination to generate molecular diversity and screening/selection to take advantage of that molecular diversity. That is, a family of nucleic acid molecules is created showing substantial sequence and/or structural identity but differing as to the presence of mutations. These sequences are then recombined in any of the described formats so as to optimize the diversity of mutant combinations represented in the resulting recombined library. Typically, any resulting recombinant nucleic acids or genomes are recursively recombined for one or more cycles of recombination to increase the diversity of resulting products. After this recursive recombination procedure, the final resulting products are screened and/or selected for a desired trait or property.

Alternatively, each recombination cycle can be followed by at least one cycle of screening or selection for molecules having a desired characteristic. In this embodiment, the molecule(s) selected in one round form the starting materials for generating diversity in the next round.

The cells to be evolved can be bacteria, archaebacteria, or eukaryotic cells and can constitute a homogeneous cell line or mixed culture. Suitable cells for evolution include the bacterial and eukaryotic, cell lines commonly used in genetic engineering, protein expression, or the industrial production or conversion of proteins, enzymes, primary metabolites, secondary metabolites, fine, specialty or commodity chemicals. Suitable mammalian cells include those from, e.g., mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIHM), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean,

sugarcane, tobacco, and arabidopsis; fish, algae, fungi (penicillium, aspergillus, podospora, neurospora, saccharomyces), insect (e.g., baculo lepidoptera), yeast (picchia and saccharomyces, Schizosaccharomycespombe). Also of interest are many bacterial cell types, both gram- negative and gram-positive, such as Bacillus subtilis, B. liceniformis, B. cereus, Escherichia coli, Streptomyces, Pseudomonas, Salmonella, Actinomycetes, Lactobacillius, Acelonitcbacter, Deinococcus, and Erwinia. The complete genome sequences of E. coli and Bacillus subtilis are described by Blattner et al., Science 277, 1454- 1462 (1997); Kunst et al., Nature 390, 249-256 (1997).

6.2. Screening for improved strains

Strains showing viability in initial selections are assayed more quantitatively for improvements in the desired properties before being restochastic &/or non-stochastically mutagenized with other strains.

Progeny resulting from mutagenesis of a strain, or those pre-selected for their ethanol tolerance and/or thermostability, can be plated on non-selective agar. Colonies can be picked robotically into microtiter dishes and grown. Cultures are replicated to fresh microtiter plates, and the replicates are incubated under the appropriate stress condition(s). The growth or metabolic activity of individual clones may be monitored and ranked. Indicators of viability can range from the size of growing colonies on solid media, density of growing cultures, or color change of a metabolic activity indicator added to liquid media. Strains that show the greatest viability are then mixed and stochastic &/or non-stochastically mutagenized, and the resulting progeny are rescreened under more stringent conditions

6.3. Methods for whole genome stochastic &/or non-stochastically mutagenizing by blind family stochastic &/or non-stochastically mutagenizing of parsed genomes and recursive cycles of forced integration and excision by homologous recombination, and screening for improved phenotypes

In vitro methods have been developed to stochastic &/or non-stochastically mutagenize single genes and operons, as set forth, e.g., herein. "Family" stochastic &/or non-stochastically mutagenizing of homologous genes within species and from different species is also an effective methods for accelerating molecular evolution. This section describes additional methods for extending these methods such that they can be applied to whole genomes.

In some cases, the genes that encode rate-limiting steps in a biochemical process, or that contribute to a phenotype of interest are known. This method can be used to target family stochastic &/or non-stochastically mutagenized libraries to such loci, generating libraries of organisms with high quality family stochastic &/or non-stochastically mutagenized libraries of alleles at the locus of interest. An example of such a gene would be the evolution of a host chaperonin to more efficiently chaperone the folding of an overexpressed protein in *E. coli*.

The goals of this process are to stochastic &/or non-stochastically mutagenize homologous genes from two or more species and to then integrate the stochastic &/or non-stochastic mutagenized genes into the chromosome of a target organism.

Integration of multiple stochastic &/or non-stochastic mutagenized genes at multiple loci can be achieved using recursive cycles of integration (generating duplications), excision (leaving the improved allele in the chromosome) and transfer of additional evolved genes by serially applying the same procedure.

In the first step, genes to be stochastic &/or non-stochastically mutagenized into suitable bacterial vectors are subcloned. These vectors can be plasmids, cosmids, BACS or the like. Thus, fragments from 100 bp to 100 kb can be handled. Homologous fragments are then "family stochastic &/or non-stochastically mutagenized" together (i.e. homologous

fragments from different species or chromosomal locations are homologously recombined). As a simple case, homologs from two species (say, *E. coli* and *Salmonella*) are cloned, family stochastically &/or non-stochastically mutagenized in vitro and cloned into an allele replacement vector (e.g., a vector with a positively selectable marker, a negatively selectable marker and conditionally active origin of replication). The basic strategy for whole genome family stochastically &/or non-stochastically mutagenizing of parsed (subcloned) genomes is additionally set forth in.

The vectors are transfected into *E. coli* and selected, e.g., for drug resistance. Most drug resistant cells should arise by homologous recombination between a family stochastically &/or non-stochastically mutagenized insert and a chromosomal copy of the cloned insert. Colonies with improved phenotype are screened (e.g., by mass spectroscopy for enzyme activity or small molecule production, or a chromogenic screen, or the like, depending on the phenotype to be assayed). Negative selection (i.e. suc selection) is imposed to force excision of tandem duplication. Roughly half C₀ of the colonies should retain the improved phenotype. Importantly, this process regenerates a "clean" chromosome in which the wild type locus is replaced with a family stochastically &/or non-stochastically mutagenized fragment that encodes a beneficial allele. Since the chromosome is "clean" (i.e., has no vector sequences), other improved alleles can also be moved into this point on the chromosome by homologous recombination.

Selection or screening for improved phenotype can occur either after step 3 or step 4. If selection or screening takes place after step 3, then the improved allele can be conveniently moved to other strains by, for example, P1 transduction. One can then regenerate a strain containing the improved allele but lacking vector sequences by "negative selection" against the suc marker. In subsequent rounds, independently identified improved variants of the gene can be sequentially moved into the improved strain (e.g., by P1 transduction of the drug marked tandem duplication above). Transductants are screened for further improvement in phenotype by virtue of receiving the transduced tandem duplication, which itself contains the family stochastically &/or non-stochastically mutagenized genetic material. Negative selection is again imposed and the process of stochastically &/or non-stochastically mutagenizing the improved strain is recursively repeated as desired.

Although this process was described with reference to targeting a gene or genes of interest, it can be used "blindly," making no assumptions about which locus is to be targeted. For example, the whole genome of an organism of interest is cloned into manageable fragments (e.g., 10 kb for plasmid-based methods). Homologous fragments are then isolated from related species. Forced recombination with chromosomal homologs creates chimeras.

6.4. Selection and screening

Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker and then physically separates the cells having the desired property. Selection is a form of screening in which identification and physical separation are achieved simultaneously, for example, by expression of a selectable marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include, for example, luciferase, beta-galactosidase, and green fluorescent protein.

Screening can also be done by observing such aspects of growth as colony size, halo formation, etc. Additionally, screening for production of a desired compound, such as a therapeutic drug or "designer chemical" can be accomplished by observing binding of cell products to a receptor or ligand, such as on a solid support or on a column. Such screening can additionally be accomplished by binding to antibodies, as in an ELISA. In some instances the screening process is preferably automated so as to allow screening of suitable numbers of colonies or cells. Some examples of automated screening devices include fluorescence activated cell sorting, especially in conjunction with cells immobilized in agarose (see Powell et. al. *Bio/Technology* 8:333-337 (1990); Weaver et. al. *Methods* 2:234- 247 (1991)), automated ELISA assays, scintillation proximity assays (Hart, H.E. et al., *Molecular Immunol.* 16:265-267 (1979)) and the formation of fluorescent, coloured or UV absorbing compounds on agar plates or in microtitre wells (Krawiec, S., *Devel. Indust. Microbiology* 31:103-114 (1990)).

Selectable markers can include, for example, drug, toxin resistance, or nutrient synthesis genes. Selection is also done by such techniques as growth on a toxic substrate to select for hosts having the ability to detoxify a substrate, growth on a new nutrient source to select for hosts having the ability to utilize that nutrient source, competitive growth in culture based on ability to utilize a nutrient source, etc.

In particular, uncloned but differentially expressed proteins (e.g., those induced in response to new compounds, such as biodegradable pollutants in the medium) can be screened by differential display (Appleyard et al. *Mol. Gen. Gent.* 247:338-342 (1995)).

Hopwood (Phil Trans R. Soc. Lond B 324:549-562) provides a review of screens for antibiotic production. Omura (Microbio. Rev. 50:259-279 (1986) and Nisbet (Ann Rev. Med. Chem. 21:149-157 (1986)) disclose screens for antimicrobial agents, including supersensitive bacteria, detection of beta-lactamase and D,D- carboxypeptidase inhibition, beta-lactamase induction, chromogenic substrates and monoclonal antibody screens.

Antibiotic targets can also be used as screening targets in high throughput screening. Antifungals are typically screened by inhibition of fungal growth. Pharmacological agents can be identified as enzyme inhibitors using plates containing the enzyme and a chromogenic substrate, or by automated receptor assays. Hydrolytic enzymes (e.g., proteases, amylases) can be screened by including the substrate in an agar plate and scoring for a hydrolytic clear zone or by using a colorimetric indicator (Steele et al. Ann. Rev. Microbiol. 45:89-106 (1991)). This can be coupled with the use of stains to detect the effects of enzyme action (such as congo red to detect the extent of degradation of celluloses and hemicelluloses).

Tagged substrates can also be used. For example, lipases and esterases can be screened using different lengths of fatty acids linked to umbelliferyl. The action of lipases or esterases removes this tag from the fatty acid, resulting in a quenching or enhancement of umbelliferyl fluorescence. These enzymes can be screened in microtiter plates by a robotic device.

6.4.1. FACS

Fluorescence activated cell sorting (FACS) methods are also a powerful tool for selection/screening. In some instances a fluorescent molecule is made within a cell (e.g., green fluorescent protein). The cells producing the protein can simply be sorted by FACS. Gel microdrop technology allows screening of cells encapsulated in agarose microdrops (Weaver et al. Methods 2:234-247 (1991)). In this technique products secreted by the cell (such as antibodies or antigens) are immobilized with the cell that generated them. Sorting and collection of the drops containing the desired product thus also collects the cells that made the product, and provides a ready source for the cloning of the genes encoding the

desired functions. Desired products can be detected by incubating the encapsulated cells with fluorescent antibodies (Powell et al. *Bio/Technology* 8:333-337 (1990)). FACS sorting can also be used by this technique to assay resistance to toxic compounds and antibiotics by selecting droplets that contain multiple cells (i.e., the product of continued division in the presence of a cytotoxic compound; Goguen et al. *Nature* 363:189-190 (1995)). This method can select for any enzyme that can change the fluorescence of a substrate that can be immobilized in the agarose droplet.

6.4.2. Reporter molecule

In some embodiments of the invention, screening can be accomplished by assaying reactivity with a reporter molecule reactive with a desired feature of, for example, a gene product. Thus, specific functionalities such as antigenic domains can be screened with antibodies specific for those determinants.

6.4.3. Cell-cell indicator

In other embodiments of the invention, screening is preferably done with a cell-cell indicator assay. In this assay format, separate library cells (Cell A, the cell being assayed) and reporter cells (Cell B, the assay cell) are used.

Only one component of the system, the library cells, is allowed to evolve. The screening is generally carried out in a two-dimensional immobilized format, such as on plates. The products of the metabolic pathways encoded by these genes (in this case, usually secondary metabolites such as antibiotics, polyketides, carotenoids, etc.) diffuse out of the library cell to the reporter cell. The product of the library cell may affect the reporter cell in one of a number of ways.

The assay system (indicator cell) can have a simple readout (e.g., green fluorescent protein, luciferase, beta-galactosidase) which is induced by the library cell product but which does not affect the library cell. In these examples the desired product can be detected by colorimetric changes in the reporter cells adjacent to the library cell.

6.4.4. Feedback mechanism

In other embodiments, indicator cells can in turn produce something that modifies the growth rate of the library cells via a feedback mechanism. Growth rate feedback can detect and accumulate very small differences. For example, if the library and reporter cells are competing for nutrients, library cells producing compounds to inhibit the growth of the reporter cells will have more available nutrients, and thus will have more opportunity for growth. This is a useful screen for antibiotics or a library of polyketide synthesis gene clusters where each of the library cells is expressing and exporting a different polyketide gene product.

6.4.5. Secretion

Another variation of this theme is that the reporter cell for an antibiotic selection can itself secrete a toxin or antibiotic that inhibits growth of the library cell. Production by the library cell of an antibiotic that is able to suppress growth of the reporter cell will thus allow uninhibited growth of the library cell.

Conversely, if the library is being screened for production of a compound that stimulates the growth of the reporter cell (for example, in improving chemical syntheses, the library cell may supply nutrients such as amino acids to an auxotrophic reporter, or growth factors to a growth-factor-dependent reporter. The reporter cell in turn should produce a compound that stimulates the growth of the library cell. Interleukins, growth factors, and nutrients are possibilities. Further possibilities include competition based on ability to kill surrounding cells, positive feedback loops in which the desired product made by the evolved cell stimulates the indicator cell to produce a positive growth factor for cell A, thus indirectly selecting for increased product formation.

In some embodiments of the invention it can be advantageous to use a different organism (or genetic background) for screening than the one that will be used in the final product. For example, markers can be added to DNA constructs used for recursive sequence

recombination to make the microorganism dependent on the constructs during the improvement process, even though those markers may be undesirable in the final recombinant microorganism.

Likewise, in some embodiments it is advantageous to use a different substrate for screening an evolved enzyme than the one that will be used in the final product. For example, Evnin et al. (Proc. Natl. Acad. Sci. U.S.A. 87:6659-6663 (1990)) selected trypsin variants with altered substrate specificity by requiring that variant trypsin generate an essential amino acid for an arginine auxotroph by cleaving arginine beta-naphthylamide. This is thus a selection for arginine-specific trypsin, with the growth rate of the host being proportional to that of the enzyme activity.

The pool of cells surviving screening and/or selection is enriched for recombinant genes conferring the desired phenotype (e.g. altered substrate specificity, altered biosynthetic ability, etc.). Further enrichment can be obtained, if desired, by performing a second round of screening and/or selection without generating additional diversity.

The recombinant gene or pool of such genes surviving one round of screening/selection forms one or more of the substrates for a second round of recombination. Again, recombination can be performed in vivo or in vitro by any of the recursive sequence recombination formats described above.

If recursive sequence recombination is performed in vitro, the recombinant gene or genes to form the substrate for recombination should be extracted from the cells in which screening/selection was performed. Optionally, a subsequence of such gene or genes can be excised for more targeted subsequent recombination. If the recombinant gene(s) are contained within episomes, their isolation presents no difficulties. If the recombinant genes are chromosomally integrated, they can be isolated by amplification primed from known sequences flanking the regions in which recombination has occurred. Alternatively, whole genomic DNA can be isolated, optionally amplified, and used as the substrate for recombination. Small samples of genomic DNA can be amplified by whole genome amplification with degenerate primers (Barrett et al. Nucleic Acids Research 23:3488-

3492 (1995)). These primers result in a large amount of random 3' ends, which can undergo homologous recombination when reintroduced into cells.

If the second round of recombination is to be performed in vivo, as is often the case, it can be performed in the cell surviving screening/selection, or the recombinant genes can be transferred to another cell type (e.g., a cell is type having a high frequency of mutation and/or recombination). In this situation, recombination can be effected by introducing additional DNA segment(s) into cells bearing the recombinant genes. In other methods, the cells can be induced to exchange genetic information with each other by, for example, electroporation. In some methods, the second round of recombination is performed by dividing a pool of cells surviving screening/selection in the first round into two subpopulations. DNA from one subpopulation is isolated and transfected into the other population, where the recombinant gene(s) from the two subpopulations recombine to form a further library of recombinant genes. In these methods, it is not necessary to isolate particular genes from the first subpopulation or to take steps to avoid random shearing of DNA during extraction. Rather, the whole genome of DNA sheared or otherwise cleaved into manageable sized fragments is transfected into the second subpopulation. This approach is particularly useful when several genes are being evolved simultaneously and/or the location and identity of such genes within chromosome are not known.

The second round of recombination is sometimes performed exclusively among the recombinant molecules surviving selection. However, in other embodiments, additional substrates can be introduced. The additional substrates can be of the same form as the substrates used in the first round of recombination, i.e., additional natural or induced mutants of the gene or cluster of genes, forming the substrates for the first round. Alternatively, the additional substrate(s) in the second round of recombination can be exactly the same as the substrate(s) in the first round of replication.

After the second round of recombination, recombinant genes conferring the desired phenotype are again selected. The selection process proceeds essentially as before. If a suicide vector bearing a selective marker was used in the first round of selection, the same vector can be used again. Again, a cell or pool of cells surviving selection is selected. If a pool of cells, the cells can be subject to further enrichment.

6.5. Factors involved in discovery and development of new drugs

In the discovery and development of new drugs, it is a common strategy to first try to identify molecules or complexes of molecules, naturally occurring within cells, that are involved in producing symptoms of a disease. These naturally occurring molecules can be thought of as "targets." A second major part of the strategy is then to find molecules that bind to the targets. These molecules are candidates for drug development, on the theory that a molecule that binds to a target can modulate (inhibit or enhance) the function of the target, thereby causing a change in the biological status of the cell containing the target. The change caused in the cell (e.g., a change in phenotype towards wild type, or a change in growth rate) may be therapeutically beneficial to the animal or human host of the cell.

The genomics revolution, by determining the DNA sequences of great numbers of genes from many different organisms, has considerably broadened the possibilities for drug discovery by identifying, large numbers of molecules that are potential targets of drug action. These technical advances in genomics however, have posed an entirely new set of challenges. Specifically, how can one prove that a chosen target molecule is essential to maintaining the disease or disorder to be treated? That is, how does one validate a target? Although methods currently available to validate targets do provide some guidelines in selection of drug targets, they are usually not conducted under the conditions in which a drug actually interacts with its target, and therefore provide a limited set of information. In addition, they do not directly address, among other things: 1) if a wild type (normal) target is essential for cell growth and viability during the disease state; 2) if the wild type gene products themselves are suitable targets for drug discovery; 3) if specific sites on a target are suitable for drug interaction (for example, in a pathogenic organism, there can be one gene coding for a single protein target with two activities -- one activity essential for growth and infectivity, the second activity non-essential); 4) if a compensatory mechanism in the cell, either in vitro or in vivo, can overcome or compensate for target modulation or, 5) if a disease state can be cured by modulation of function of the candidate target. These methods also do not provide a direct route for testing wild type target proteins in high throughput screening assays.

An analysis of the discovery of novel antimicrobial agents illustrates the problems researchers in all fields of drug development face today. The increasing prevalence of drug-resistant pathogens (bacteria, fungi, parasites, etc.) has led to significantly higher mortality rates from infectious diseases and currently presents a serious crisis worldwide. Despite the introduction of second and third generation antimicrobial drugs, certain pathogens, such as vancomycin resistant strains of *Enterococcus faecium*, have developed resistance to all currently available drugs.

New antimicrobial drugs must be discovered to treat such infections by such organisms, and new methods are urgently needed to facilitate making such discoveries.

Neither whole cell screening, chemistry nor target based drug discovery approaches as currently applied, have met the challenge of controlling infectious diseases, particularly those caused by drug resistant microorganisms. Whole cell screening assays have been limited by the fact that they are unable to identify compounds that can effectively modulate a target function inside the cell but cannot permeate the cell membrane to get to the target. Therefore entire classes of potent, intracellular target modulators, which could be subsequently modified by medicinal chemistry to increase cell membrane permeability, go undetected. Chemistry based approaches have focused on chemically modifying the molecular structure of existing antimicrobial drugs or combining existing antimicrobials with another agent to circumvent established resistance mechanisms. Technical advances in molecular biology, automated methods for high throughput screening and chemical syntheses have led to an increase in the number of target based screens utilized for antimicrobial drug discovery and in the number of compounds being analyzed. However, despite these advances, only a limited number of antimicrobial drugs acting by a novel mechanism have been identified during recent years.

How does one efficiently establish screening assays for drugs that can be used with a variety of different targets having different properties, enzymatic activities, or even unknown functions? A number of potentially novel, valuable targets are incompatible with current methods to screen for drug candidates because either the target's exact function and molecular mechanism of action are unknown, or there are technical obstacles preventing the development of effective high throughput screening methods. It can take anywhere

from six months to several years to develop a screening assay, which is impractical when the goal is to rapidly screen multiple targets in a cost-effective manner.

The path in the progression from target identification through assay development, high throughput screening, medicinal chemistry, lead optimization, preclinical and clinical drug development is expensive, time consuming and full of technical challenges. Many different targets must be screened against multiple chemical compounds to identify new lead compounds for drug development. New, efficient technologies are needed that can be broadly applied to a variety of different targets to validate targets in the direct context of the desired outcome of drug therapy and to rapidly develop screening assays using these targets for drug discovery. Such developments will allow the wealth of genomics information to be leveraged for drug discovery and will lower the risk and costs while expediting the timelines of the drug discovery process.

6.6. General approach for identifying targets

The invention relates to methods that couple the validation of a target for drug discovery with the development of an assay to identify compounds that cause a phenotypic effect on the target cell. These procedures can be applied to identifying compounds that bind to and modulate the function of target components of a cell whose function is known or unknown, and cell components that are not amenable to other screening methods.

The invention relates to procedures for identifying a compound that binds to and modulates (inhibits or enhances) the function of a component of a cell, thereby producing a phenotypic effect in the cell. Within these procedures are methods for identifying a biomolecule that 1) binds to, in vitro, a component of a cell that has been isolated from other constituents of the cell and that 2) causes, in vivo, as seen in an assay upon intracellular expression of the biomolecule, a phenotypic effect in the cell which is the usual producer and host of the target cell component. In an assay demonstrating characteristic 2) above, intracellular production of the biomolecule can be in cells grown in culture or in cells introduced into an animal. Further methods within these procedures are those methods comprising an assay for a phenotypic effect in the cell upon intracellular production of the biomolecule, either in cells in culture or in cells that have been introduced into one or more animals, and an assay to identify one or more compounds that behave as competitors of the biomolecule in an assay of binding to the target cell component.

6.6.1. Process for identifying one or more compounds that produce a phenotypic effect on a cell

One procedure envisioned in the invention is a process for identifying one or more compounds that produce a phenotypic effect on a cell. The process is at the same time a method for target validation. The process is characterized by identifying a biomolecule which binds an isolated target cell component, constructing cells comprising the target cell component and further comprising a gene encoding the biomolecular binder which can be expressed to produce the biomolecular binder, testing the constructed cells for their ability

to produce, upon expression of the gene encoding the biomolecular binder, a phenotypic effect in the cells (e.g., inhibition of growth), wherein the test of the constructed cells can be a test of the cells in culture or a test of the cells after introducing them into host animals, or both, and further, identifying, for a biomolecular binder that caused the phenotypic effect, one or more compounds that compete with the biomolecular binder for binding to the target cell component.

A test of the constructed cells after introducing them into host animals is especially well-suited to assessing whether a biomolecular binder can produce a particular phenotype by the expression (regulatable by the researcher) of a gene encoding the biomolecular binder. In this method, cells are constructed which have a gene encoding the biomolecular binder, and wherein the biomolecular binder can be produced by regulation of expression of the gene. The constructed cells are introduced into a set of animals. Expression of the gene encoding the biomolecular binder is regulated in one group of the animals (test animals) such that the biomolecular binder is produced. In another group of animals, the gene encoding the biomolecular binder is regulated such that the biomolecular binder is not produced (control animals). The cells in the two groups of animals are monitored for a phenotypic change (for example, a change in growth rate). If the phenotypic change is observed in cells in the test animals and not in the cells in the control animals, or to a lesser extent in the control animals, then the biomolecular binder has been proven to be effective in binding to its target cell component under in vivo conditions.

A further embodiment of the invention is a method for determining whether a target cell component of a particular cell type (a "first cell") is essential to producing a phenotypic effect on the first cell, the method having the steps: isolating the target component of the first cell; identifying a biomolecular binder of the isolated target component of the first cell; constructing a second type of cells ("second cell") comprising the target component and a regulable, exogenous gene encoding the biomolecular binder; and testing the second cell in culture for an altered phenotypic effect, upon production of the biomolecular binder in the second cell; whereby, if the second cell shows the altered phenotypic effect upon production of the biomolecular binder, then the target component of the first cell is essential to producing the phenotypic effect on the first cell. The target cell component in this embodiment and in other embodiments not limited

to pathogens can be one that is found in mammalian cells, especially cells of a type found to cause or contribute to disease or the symptoms of disease (e.g., cells of tumors or cells of other types of hyperproliferative disorders).

6.6.1.1. Procedure for identifying and/or designing compounds with antimicrobial activity against a pathogen

The invention further relates to methods particularly well suited to a procedure for identifying and/or designing compounds with antimicrobial activity against a pathogen whose target cell component is the subject of studies to identify such compounds. A common mechanism of action of an antimicrobial agent is binding to a component of the cells of the pathogen treated with the antimicrobial.

The procedure includes methods for identifying biomolecules that bind to a chosen target in vitro, methods for identifying biomolecules that also bind to the chosen target and modulate its function during infection of a host mammal in vivo, and methods for identifying compounds that compete with the biomolecules for sites on the target in competitive binding assays. Compounds identified by this procedure are candidates for drugs with antimicrobial activity against the pathogen.

6.6.1.1.1. Identifying a biomolecular inhibitor of growth of pathogen cells

One embodiment of the invention is a method for identifying a biomolecular inhibitor of growth of pathogen cells by using cell culture techniques, comprising contacting one or more types of biomolecules with isolated target cell component of the pathogen, applying a means of detecting bound complexes of biomolecules and target cell component, whereby, if the bound complexes are detected, one or more types of biomolecules have been identified as a biomolecular binder of the target cell component, constructing a pathogen strain having a regulable gene encoding the biomolecular binder, regulating expression of the gene encoding the biomolecular binder to express the gene; and monitoring growth of the pathogen cells in culture relative to suitable control cells, whereby, if growth of the pathogen cells is decreased compared to growth of suitable

control cells, then the biomolecule is a biomolecular inhibitor of growth of the pathogen cells.

6.6.1.1.2. Identifying compounds that inhibit infection of a mammal by a pathogen

A further embodiment of the invention is a method, employing an animal test, for identifying one or more compounds that inhibit infection of a mammal by a pathogen by binding to a target cell component, comprising constructing a pathogen comprising a regulable gene encoding a biomolecule which binds to the target cell component, infecting test animals with the pathogen, regulating expression of the regulable gene to produce the biomolecule, monitoring the test animals and suitable control animals for signs of infection, wherein observing fewer or less severe signs of infection in the test animals than in suitable control animals indicates that the biomolecule is a biomolecular inhibitor of infection, and identifying one or more compounds that compete with the biomolecular inhibitor of growth for binding to the target cell component (as by employing a competitive binding assay), then the compound inhibits infection of a mammal by a pathogen by binding to a target.

The competitive binding assay to identify binding analogs of biomolecular binders, which have been proven to bind to their targets in an intracellular test of binding, can be applied to any target for which a biomolecular binder has been identified, including targets whose function is unknown or targets for which other types of assays are not easily developed and performed. Therefore, the method of the invention offers the advantage of decreasing assay development time when using a gene product of known function as a target cell component and the advantage of bypassing the major hurdle of gene function identification when using a gene product of unknown function as a target cell component.

Other embodiments of the invention are cells comprising a biomolecule and a target cell component, wherein the biomolecule is produced by expression of a regulable gene, and wherein the biomolecule modulates function of the target cell component, thereby causing a phenotypic change in the cells. Yet other embodiments are cells comprising a biomolecule and a target cell component, wherein the biomolecule is a biomolecular binder of the target cell component, and is encoded by a regulable gene. The cells can

include mammalian cells or cells of a pathogen, for instance, and the phenotypic change can be a change in growth rate.

The pathogen can be a species of bacteria, yeast, fungus, or parasite, for example.

6.6.2. Definitions

Target: (also, "target component of a cell," or "target cell component") a constituent of a cell which contributes to and is necessary for the production or maintenance of a phenotype of the cell in which it is found. A target can be a single type of molecule or can be a complex of molecules. A target can be the product of a single gene, but can also be a complex comprising more than one gene product (for example, an enzyme comprising alpha and beta subunits, mRNA, tRNA, ribosomal RNA or a ribonucleoprotein particle such as a snRNP). Targets can be the product of a characterized gene (gene of known function) or the product of an uncharacterized gene (gene of unknown function).

Target Validation: the process of determining whether a target is essential to the maintenance of a phenotype of the cell type in which the target normally occurs. For example, for pathogenic bacteria, researchers developing antimicrobials want to know if a compound which is potentially an antimicrobial agent not only binds to a target in vitro, but also binds to, and modulates the function of, a target in the bacteria in vivo, and especially under the conditions in which the bacteria are producing an infection — those conditions under which the antimicrobial agent must work to inhibit bacterial growth in an infected animal or human. If such compounds can be found that bind to a target in vitro and alter the target's function in cells resulting in an altered phenotype, as found by testing cells in culture and/or as found by testing cells in an animal, then the target is validated.

Phenotypic Effect: a change in an observable characteristic of a cell which can include, e.g., growth rate, level or activity of an enzyme produced by the cell, sensitivity to various agents, antigenic characteristics, and level of various metabolites of the cell. A phenotypic effect can be a change away from wild type (normal) phenotype, or can be a change towards wild type phenotype, for example.

A phenotypic effect can be the causing or curing of a disease state, especially where mammalian cells are referred to herein. For cells of a pathogen or tumor cells, especially, a phenotypic effect can be the slowing of growth rate or cessation of growth.

Biomolecule: a molecule which can be produced as a gene product in cells that have been appropriately constructed to comprise one or more genes encoding the biomolecule. Preferably, production of the biomolecule can be turned on, when desired, by an inducible promoter. A biomolecule can be a peptide, polypeptide, or an RNA or RNA oligonucleotide, a DNA or DNA oligonucleotide, but is preferably a peptide. The same biomolecules can also be made synthetically. For peptides, see Merrifield, J., J. Am. Chem. Soc. 85: 2140-2154 (1963). For instance, an Applied Biosystems 431 A Peptide Synthesizer (Perkin Elmer) can be used for peptide synthesis. Biomolecules produced as gene products intracellularly are tested for their interaction with a target in the intracellular steps described herein (tests performed with cells in culture and tests performed with cells that have been introduced into animals). The same biomolecules produced synthetically are tested for their binding to an isolated target in an initial in vitro method described herein.

Synthetically produced biomolecules can also be used for a final step of the method for finding compounds that are competitive binders of the target.

Biomolecular Binder (of a target): a biomolecule which has been tested for its ability to bind to an isolated target cell component in vitro and has been found to bind to the target.

Biomolecular Inhibitor of Growth: a biomolecule which has been tested for its ability to inhibit the growth of cells constructed to produce the biomolecule in an "in culture" test of the effect of the biomolecule on growth of the cells, and has been found, in fact, to inhibit the growth of the cells in this test in culture.

Biomolecular Inhibitor of Infection: a biomolecule which has been tested for its ability to ameliorate the effects of infection, and has been found to do so. In the test, pathogen cells constructed to regulably express the biomolecule are introduced into one or more animals, the gene encoding the biomolecule is regulated so as to allow production of the

biomolecule in the cells, and the effects of production of the biomolecule are observed in the infected animals compared to one or more suitable control animals.

Isolated: term used herein to indicate that the material in question exists in a physical milieu distinct from that in which it occurs in nature. For example, an isolated target cell component of the invention may be substantially isolated with respect to the complex cellular milieu in which it naturally occurs. The absolute level of purity is not critical, and those skilled in the art can readily determine appropriate levels of purity according to the use to which the material is to be put.

In many circumstances the isolated material will form part of a composition (for example, a more or less crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography (for example, HPLC).

Pathogen or Pathogenic Organism: an organism which is capable of causing disease, detectable by signs of infection or symptoms characteristic of disease. Pathogens can include procaryotes (which include, for example, medically significant Gram- positive bacteria such as *Streptococcus pneumoniae*, *Enterococcus faecalis* and *Staphylococcus aureus*, Gram-negative bacteria such as *Escherichia coli*, *Pseudomonas aeruginosa* and *Klebsiella pneumoniae*, and "acid-fast" bacteria such as *Mycobacteria*, especially *M. tuberculosis*), eucaryotes such as yeast and fungi (for example, *Candida albicans* and *Aspergillus fumigatus*) and parasites. It should be recognized that pathogens can include such organisms as soil-dwelling organisms and "normal flora" of the skin, gut and orifices, if such organisms colonize and cause symptoms of infection in a human or other mammal, by abnormal proliferation or by growth at a site from which the organism cannot usually be cultured.

6.7. Target validation

The present invention relates to methods that couple the validation of a target cell component for drug discovery with the development of a validated assay to identify compounds that cause a phenotypic effect on the target cell (cell harboring the target cell component). When the target cells are cells of a pathogenic organism, compounds identified by this procedure are candidates for drugs with antimicrobial activity against the pathogen.

The method utilized for target validation provides a test of how a biomolecule produced intracellularly binds to a specific site on a target cell component and alters the target's function in a cell during an established infection or disease. The technology to validate the target identifies a biomolecule specific to the target that can be used in a screening assay to identify drug leads, thereby coupling target validation with drug lead identification. The method also validates specific sites on a target molecule for drug discovery, which is especially important for proteins involved in multiple functions.

6.7.1. Intracellular validation of a biomolecule

Described herein are methods that result in the identification of compounds that cause a phenotypic effect on a cell. The general steps described herein to find a compound for drug development can be thought of as these: (1) identifying a biomolecule that can bind to an isolated target cell component in vitro, (2) confirming that the biomolecule, when produced in cells with the target cell component, can cause a desired phenotypic effect and (3) identifying, by an in vitro screening method, for example, compounds that compete with the biomolecule for binding to the target cell component. Advantages of these steps are that it is not necessary to identify the function of the target cell component and it is not necessary to develop an assay tailored to the function (e.g., enzyme activity) of the target cell component.

Central to these methods is general step (2) above, intracellular validation of a biomolecule comprising one or more steps that determine whether a biomolecule can cause a phenotypic effect on a cell, when the biomolecule is produced by the expression